PB82-226432

Application of Cluster Analysis to Aerometric Data. Volume I. Part 1: Clustering Validation and Classification of Data. Part 2 Investigation and Report of Cluster Anaysis

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

Dec 80

U.S. Department of Commerce
National Technical Information Service
NTIS

TYPING GUIDE SHEET
BEGIN FIRST LINE OF TEXT HERE
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD; BEGIN SECTIONS HERE
6-1/2''
9-1/8''

PB82-226432

EPA-600/4-80-048A
December 1980

APPLICATION OF CLUSTER ANALYSIS
TO
AEROMETRIC DATA

Volume I

Edited by
Raymond C. Rhodes
Quality Assurance Specialist

Part 1: Clustering, Validation, and Classification of Data
by Harold Crutcher, Consultant
Asheville, North Carolina

Part 2: Investigation and Report of Cluster Analysis
by PEDCO Environmental, Inc.
Durham, North Carolina

December 1980

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8''
PAGE NUMBER

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

EPA-287 (Cin.)
(4-76)

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO.
EPA-600/4-80-048-a

2.

3. RECIPIENT'S ACCESSION NO.
PB82-226432

4. TITLE AND SUBTITLE Application of Cluster Analysis to Aerometric Data, Volume I. Part 1: Clustering, Validation, and Classification of Data. Part 2: Investigation and Report of Cluster Analysis

5. REPORT DATE
December 1980

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S) Editor: Raymond C. Rhodes
Part 1: Harold L. Crutcher
Part 2: Carl Nelson, B. Fairbairn, and M. Symons

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Environmental Protection Agency
Environmental Monitoring Systems Laboratory
Research Triangle Park, NC 27711

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE
EPA 600/08

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The calibration and enhancement of Wolfe's NORMIX (normal mixtures) computer program in the National Computing Center of the U.S. Environmental Protection Agency at the Research Triangle Park, NC is documented. The program is available for data clustering, validation, and classification.

In Part 1, an historical perspective of the NORMIX algorithm is presented with details of calibration, using the Fisher-Anderson Iris data set. New program options are then introduced with illustrative outputs including clustering features such as variate profiles and bivariate representations.

In Part 2, SAS and MIKCA clustering algorithms are applied to LACS data to investigate their performance as a multivariate tool for modeling air pollution data.

In Part 3, a discussion of new NORMIX modifications and options is provided with complete program listings and other user aids.

In Part 4, tabular and graphical displays record some aspects of the separation of environmenal data into clusters by the NORMIX program. Three CHAMP monitoring sites are principal data sources.

This volume, Volume III, of a 3-volume series comprises the report for Part 4. Parts 1 and 2 are reported in Volume I; Part 3 is reported in Volume II.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Cluster Analysis Data Validation Data Classification Outliers Statistics Environmental Data	Environmental Monitoring Data Management	43F 68A

18. DISTRIBUTION STATEMENT
RELEASE TO PUBLIC

19. SECURITY CLASS *(This Report)*
UNCLASSIFIED

20. SECURITY CLASS *(This page)*
UNCLASSIFIED

21. NO. OF PAGES
143

22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

PREFACE

Two separate studies have been pursued in the application of multivariate cluster analysis to environmental air monitoring data. The first study, which appears in this volume as Part 2, recommended the second study. The second study is reported in three parts, namely, Part 1 of this volume and Parts 3 and 4, which are respectively Volumes II and III of the complete report. Volumes II and III are available under separate covers. Part 1 contains an extensive historical perspective. For this reason, Part 1 (the first part of the second study) and Part 2 (the first study) are contained in this volume.

Previous uses of cluster analyses of air pollution data have primarily concerned the identification of specific pollutant sources or types of sources which result in particular combinations of pollutants. Although such applications will continue to be useful for these purposes, it was considered desirable to explore a more general application of multivariate cluster analysis.

In most summarizations and analyses of air monitoring data, the data of each of several concomitant measurements are considered separately as univariate distributions. In actuality, the data should more correctly be considered for some purposes as multivariate observations which involve (a) correlations among various pollutants, (b) correlations between and among pollutants and meteorological variates, (c) cyclic patterns (diurnal, day-of-week, and seasonal), and (d) local source effects. All of the above relationships are reflected in the different combinations of aerometric variates which cluster together. Because of these clustering tendencies of aerometric data, it is believed that multivariate cluster analysis or classification techniques may be fruitful either (a) as preliminary analyses to application of other techniques, or (b) as sufficiently revealing on their own.

The study presented here first but chronologicaly second, involved the use of the NORMIX (NORMAL MIXTURES) computer program. The NORMIX program has a substantial statistical basis with corresponding statistical assumptions

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

and significance tests. The program was used in the study of atmospheric pollutants and simultaneous meteorological information at three sites: Angwin, CA; Loma Linda, CA; and Magna, UT. These sites were involved in the Health Effects Research Laboratory, Community Health Air Monitoring Program (CHAMP). The usefulness of the data validation potential of the program was demonstrated in the initial processing of the data. Outliers--singleton, doubleton, and in larger sets--were isolated in the processing, identified as mistakes in data preparation and corrected. With the elimination of the artificial outliers created in the preparation of the data, sets of data devoid of extreme outliers were available. However, it is emphasized that the program would isolate outliers caused by other types of errors in the data generation, as well as outlying but valid data sets which require critical technical examination as to their reality and cause.

The second project has resulted in the significant and extended development of a comprehensive NORMIX program which combines many new features and options not previously available in a single package. This computer package is completely operational on the Univac 1110 system at the National Computer Center (NCC), Research Triangle Park, N.C. Some of the options produce printer output of histogram analysis of the input variate data sets, transformations of data, and printer output of histogram analysis of variates within each cluster. An option also permits the development of data files for entry into computer plotting routines. A rather complete description and listing of the computer program is contained in Volume II of this report series (EPA-600/4-80-048-b). Volume III of this report series (EPA-600/4-80-048-c) provides examples of application of the software program package to environmental data.

The reader and user of this report will find a rather extensive discussion of clustering and data validation problems and the use of computer programs and clustering techniques. Limitations of the programs are also discussed.

Chronologically, the first of the two studies, now presented here, was an exploratory application of the two clustering algorithms, those of the Statistical Analysis Systems (SAS) and the Multivariate Iterative K-Means Cluster Analysis (MIKCA). These algorithms are essentially non-statistical in nature, and although certain statistics are computed, no statistical significance tests are used.

In these algorithms, clusters of similar data are formed and are differentiated from other clusters by selected geometric or statistical distance. Options are available to use different distance criteria. The SAS and MIKCA programs were applied to the meteorological, pollutant, and traffic count data produced at a Los Angeles Catalyst Study (LACS) site. The particular variates used in the Environmental Monitoring Systems Laboratory (EMSL) study were suspended particulates, lead, carbon monoxide, windspeed, wind direction, temperature, and traffic count. For the data compared, the two algorithms produced similar results. In addition, a multivariate regression analysis was performed for comparison purposes. A third set of algorithms was suggested, namely the NORMIX computer program, the discussion and applications of which constitutes the first part of this combined report.

As mentioned above, the study involving the NORMIX program demonstrated the effective use of the multivariate technique to detect multivariate outliers--outliers which are not identifiable as such by use of univariate techniques, but which are revealed as very unusual combinations of the variates. This suggests the use of cluster analysis as a technique for data validation purposes. Perhaps it would be too heavy a technique for use solely for data validation, but would be quite valuable as a screening technique prior to other analyses.

Data from monitoring activities cannot be considered as random samplings from some single universe, but rather result from samplings of mixtures of distributions usually internally correlated within each group. These studies showed that, as expected, pollutant data depend on, or are highly correlated with, meteorological conditions. Under similar meteorological conditions, regression analysis would be a useful technique. Weather conditions do exhibit clustering tendencies. At a given site and with a given set of pollutant sources, it is obvious that the pollutant concentrations at the site are highly dependent upon the existing meteorological

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

regime. Thus, the pollutant data distributions will be mixtures of the distributions which result from the mixture of the meteorological regimes existing over the time period of monitoring. This suggests that the pollutant data distributions first be separated by meteorological regimes (such as by use of cluster analysis) and that the distributions be evaluated individually. Analysis involving prediction of future pollutant concentrations should then consider not only the pollutant concentration distributions for each meteorological regime, but should consider in addition, the future probabilities of each occurrence of each of the meteorological regimes.

With the three computer clustering programs discussed and available at the Environmental Protection Agency, Research Triangle Park installation, the user has available tools to use in study of atmospheric pollution problems. Pollutant source locations, the identity and timing of the pollutant episodes can be identified. It is hoped that the new and comprehensive Cluster Analysis Program will be employed to further enhance the analysis, understanding, and effective use of air monitoring data.

R. C. Rhodes
Project Officer

6-1/2"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLE OF CONTENTS

Part 1 Clustering, Validation, and Classification of Data

Part 2 Investigation and Report of Cluster Analysis

6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TYPING GUIDE SHEET
BEGIN FIRST LINE OF TEXT HERE
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD; BEGIN SECTIONS HERE
6-1/2"
9-1/8"

PART I

CLUSTERING, VALIDATION AND CLASSIFICATION OF DATA

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

ACKNOWLEDGMENTS

The valuable assistance given by Mr. Raymond C. Rhodes, Data Management and Analysis Division, Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, who acted as the Project Officer of the cluster analysis project is gratefully acknowledged. My appreciation is also extended to the following individuals for providing the necessary support: Mr. Maurice E. Graves, Northrop Services, Inc., for programming options and preparing data; Mr. James A. Stephenson, Integrated Systems, Inc., for adapting NORMIX to the UNIVAC 1110 and writing ellipse-plotting routines; Ms. Carole B. Cameron, Northrop Services, Inc., for developing the graphics presentations; Ms. Debora R. Pizer and Mr. José M. Suñé, Statistical and Technical Analysis Branch, Environmental Monitoring and Support Laboratory, U.S. Environmental Protection Agency, for providing data and supplying site information. Appreciation is expressed to Ms. Olga Wierbicki, Northrop Services, Inc., for her prompt and careful editing of the manuscript.

Statements have been reproduced from B. S. Everitt "Unresolved Problems in Cluster Analysis," Biometrics 35:169-181, 1979 with the kind permission of both the Biometric Society and Dr. B. S. Everitt. The Editors of the Annals of Mathematical Statistics and Dr. W. J. Dixon have also given their kind permission to quote from the paper, "Analysis of Extreme Values." Appreciation is also tendered to Dr. John H. Wolfe to quote freely from SRM-682 of August 1967 and other works of his.

Permission has also been granted to quote from page 16, "Outliers in Statistical Data," by Dr. V. Barnett and Prof. T. Lewis, copyright 1978 by John Wiley and Sons, Ltd., reprinted 1979.

Appreciation is expressed to the several reviewers who made appropriate helpful cogent remarks and suggestions.

6-1/2"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

ABSTRACT

The primary objective of the work performed here is to make available to the environmental research community an electronic computer program to help classify and validate data. The secondary objective, three-fold in nature, is to (1) cluster, (2) classify data into groups of "within" group similarities and "between" group dissimilarities, and (3) use these clusters to validate data as well as to describe environmental characteristics.

The computer program is the NORMIX program of Wolfe. Many pre- and post-options have been added. These greatly enhance the applicability and usefulness of the program. The program, calibrated and viable, is now available in the National Computing Center of the U.S. Environmental Protection Agency at the Research Triangle Park, North Carolina.

Among several sets of data processed were (1) a set of physical and chemical characteristics of unused filters and (2) three sets of environmental data; pollutant and meteorological. Selected output illustrates some features of the program and its effectiveness. The three sets of environmental data were taken from the records of the CHAMP stations at (1) Angwin, CA, (2) Magna, UT, and (3) Loma Linda, CA.

6-1/2''
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8''
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLE OF CONTENTS

	Page
List of Tables	xi
List of Figures	xii
1.0 INTRODUCTION	1-1
1.1 Objective	1-1
1.2 Preliminary Discussion	1-1
1.3 Previous Work on Clustering	1-6
1.4 Unresolved Problems in Cluster Analysis	1-7
1.4.1 Order of Clustering, Factor Analysis, and Regression Analysis	1-7
1.4.2 Scaling	1-8
1.4.3 Stability	1-8
1.4.4 Significance Assessment	1-8
1.4.5 Linkage	1-9
1.4.6 Covariance Matrices	1-9
1.4.7 Number of Groups	1-10
1.4.8 Efficacy of Wolfe's Technique	1-10
1.5 Electronic Computer Programs Available to EPA-NCC at RTP	1-11
1.5.1 BMDP	1-12
1.5.2 CLUSTAN	1-13
1.5.3 IMSL	1-14
1.5.4 MIKCA	1-15
1.5.5 SAS	1-16
1.5.6 SPSS	1-17
2.0 DATA VALIDATION	2-1
2.1 Rogue or Outlier Tests	2-1
3.0 NORMAP-NORMIX	3-1
3.1 NORMIX Flow (modified) and Options	3-7
3.2 NORMIX Preparation	3-7
3.2.1 Input Data Displays	3-7
3.2.2 Transformed Data Displays and Preliminary Estimates	3-12
3.2.3 Preliminary Estimates (Hierarchical Grouping)	3-14

6-1/2"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Page

3.3 Clustering 3-24
3.3.1 Chi-square Tests 3-25
3.3.2 Histograms 3-25
3.3.4 Probability Level for Chi-square 3-26
4.0 NORMIX CALIBRATION AND OUTPUT 4-1
4.1 Calibration 4-1
4.1.1 Objective 4-1
4.2 Demonstration 4-1
4.2.1 Classical Data Set 4-1
4.2.2 Synthetic Data Set 4-5
4.2.3 Other Data Sets 4-5
4.3 RESULTS OF CALIBRATION 4-6
4.3.1 Output 4-6
5.0 PROCESSING OF SELECTED CHAMP DATA 5-1
5.1 Data Problem Examples 5-1
5.2 Station Location Descriptions 5-4
5.2.1 Angwin, California 5-4
5.2.2 Magna, Utah 5-6
5.2.3 Loma Linda, California 5-6
5.3 Selection of Data for Processing 5-9
5.4 Magna, Utah, Data Discussion 5-12
6.0 RESULTS AND CONCLUSIONS 6-1
7.0 REFERENCES 7-1
8.0 BIBLIOGRAPHY 8-1

9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLES

Number		Page
1	NORMIX Program with Iris Data	3-3
2	Fisher Iris data; input data and output classifications for 3 clusters, I. setosa, 1-50, I. versicolor, 51-100, and I. virginica, 101-150 with sepal length, sepal width, petal length, and petal width	3-4
3	Fisher Iris data; 150 sepal lengths distributed in equal probability classes with chi-square frequency table and test, histogram, and arithmetic probability chart	4-8
4	Three cluster output of Fisher Iris data; 150 input data, proportions, means, standard deviations, correlation matrices, and null hypothesis test	4-9
5	Fisher Iris data; I. setosa, 50 sepal lengths in equal probability class intervals, chi-square frequency table and test, histogram	4-10
6	Summary of significant clusters obtained by the NORMIX computing programs for selected periods of CHAMP data at Angwin, California, Magna, Utah, and Loma Linda, California	5-10
7	Three cluster output of NORMIX computer processing of Magna, Utah, day 3, pollutant and meteorological data	5-14
8	Intercomparison of selected pollutant data through the year at Angwin, Calif.,; Loma Linda, Calif.; and Magna, Utah.	5-16

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

FIGURES

Number		Page
1	NORMIX flow and options	3-8
2	NORMIX flow chart	3-9
3a	Fisher Iris data mapping flow, I. setosa (..., x)	3-18
3b	Fisher Iris data mapping flow, I. setosa (..., ≡)	3-19
3c	Schematic clustering diagram for 74 four-component vectors of Canton Island, 30-mb height, temperature, zonal components, and meridional components of the winds (1960-64)	3-23
4	Ellipses, 0.30 probability, 3 clusters, petal length versus petal width, Fisher Iris data	4-12
5	Ellipses, 0.30 probability, 4 clusters, petal length versus petal width, Fisher Iris data	4-13
6	Ellipses, 0.30 probability, 5 clusters, petal length versus petal width, Fisher Iris data	4-14
7	Profile plots for the total, and two cluster data for new filters, physical and chemical characteristics: (1) penetration (%), (2) tensile strength (kg), (3) thickness (mils), (4) flow (cmm), and (5) integrity (mg), and (6) pH	4-18
8	Profile plots for the total, and three cluster data for new filters, physical and chemical characteristics: (1) penetration (%), (2) tensile strength (kg), (3) thickness (mils), (4) flow (cmm), (5) integrity (mg), and (6) pH	4-19
9	CHAMP data example for Loma Linda, California, for day 103, 1977, with some indicated bad data, Station Identifier 0823	5-2
10	CHAMP data at Angwin, California, hour 1, showing key punch errors in lines 18 and 39, Station Identifier 0812	5-3
11	Map showing the location of the Angwin, California, CHAMP Station 0812	5-5
12	Map showing the location of the Magna, Utah, CHAMP Station 0541	5-7
13	Map showing the location of the Loma Linda, California, CHAMP Station 0823	5-8

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA, OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Number **Page**

14 Magna, Utah, day 3, 0.50 probability, ellipses of the west-east and south-north wind components for three cluster types. Winds from the west and south are positive 5-18

15 Magna, Utah, day 3, profile plots of pollutants and meteorological variates against a three sigma background 5-19

6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

CLUSTERING, VALIDATION, AND CLASSIFICATION OF DATA

1.0 INTRODUCTION

1.1 OBJECTIVE

As the title reveals, the objective of the work reported here was threefold: to cluster, validate, and classify data, in particular environmental data. Clustering techniques were assembled and developed that would permit both data validation and classification (prediction). The initial thrust of the work was to place a calibrated clustering and classification program in the National Computation Center (NCC) computer library facilities of the U.S. Environmental Protection Agency (EPA), located in Research Triangle Park near Raleigh, North Carolina. One such program is the Wolfe (1971b) NORMIX computer program. Selected data blocks of environmental information were then processed to check the efficacy of the data clustering, validation, and classification.

1.2 PRELIMINARY DISCUSSION

Heterogeneous mixtures in the animate and inanimate worlds can be separated into groups more or less homogeneous. For example, plants and animals can be categorized according to genus and species; snow, rain, or fractional crystallization of minerals according to physical and chemical structures. In reaching a decision on how to group, we often have to rely on our senses and subjective decisions which may or may not be logical, particularly in the animate world. In the inanimate world, the physical and chemical configurations are controlled by various laws, but ultimately, we have to decide where the divisions fall. To quote Bose and Roy (1938) (also quoted by Friedman and Rubin in 1967): "The problems of discrimination and classification are insistent in the sciences." Measurement, recording, and transmission (communication) systems are unfortunately always subject to errors of measurement, errors in the total system, and just plain mistakes; their effects become part of the record. Design engineering and control techniques can keep such errors to a minimum. Procedures are needed that will permit the establishment of characteristics that can serve as guides in deciding what is wanted, acceptable or unacceptable, correct or incorrect.

Examples of organization and bookkeeping requiring grouping and classification can be found already in the early history of mankind: in the clay tablets of Babylon, in the knotted thread configurations (quipus) of the Peruvian Incas, in the taxonomic system devised by Linnaeus, and in the writings of Aristotle. In each case patterns were revealed. Leaders in whatever field of endeavor have tended to be those who grasped the patterns around them and used them to advantage.

John Wolfe (1967) writes: "How do human beings arrive at the conceptual framework with which reality is perceived and understood? It could be argued that the process is basically a form of mixture analysis which detects the modes in the frequency of vast number of sensory patterns. The process may be an iterative one in which random initial hypotheses about means and dispersions are revised after relating them to old as well as new experiences. Unlike the process of learning to recognize and discriminate among preestablished patterns or concepts, the process of forming new abstractions is a slow and difficult one for both computers and men. The iterative process spans the entire life of an individual and extends through many generations in the development of a culture. The principles of mixture analysis may provide an insight into the nature of concept formation which will assist in programming more intelligent machines."

In the environmental fields, the wanted characteristics are those that maintain or improve the quality of life from all viewpoints, while the unwanted characteristics are those that are deleterious from any viewpoint. To state that any characteristic is or is not wanted may require substantive documentation. For example, the constituency of the atmosphere (air), lithosphere (earth), and hydrosphere (water) through which life has evolved to the present time and state basically is known. At question is the determination of the extent of 1) the deleterious effects of newly introduced constituents, or 2) the augmentation or diminution of known constituents to the point where deleterious effects occur. The amount and distribution of the atmospheric, lithospheric, and hydrospheric contaminants will need to be determined.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Inherent in the word "contaminant" is the concept of change from an accepted "status quo," though not always in an adverse sense.

Similarities and dissimilarities inhabit all areas of life. An individual is constantly confronted with the need to decide between them, and even his very life can depend on that ability. Many times the sense of where an item belongs may not be precisely definable, but still a decision has to be made. Mistakes will occur. Another individual may assess the situation differently, depending on his prior experiences or differing capabilities.

In all data validation situations, data may exist that are patently not permissible in one situation, yet may be permissible in another situation. For example, in wind distributions the recording and use of a 420 degree angle from the north is often cause for pause. Such a data point would be accepted by some and rejected by others. The former would see nothing wrong if the total turning of the wind, particularly with a change in altitude, was important. The latter would be only interested in the wind direction at the moment with no reference to earlier or later data. For the former, the so-called wrap-around function would be a model. In other situations, the cycloidal or epicycloidal wind or storm patterns might be an important feature for some individuals.

The main point to be taken from above is that data may not fit preconceived patterns. For certain purposes, these data may be considered outliers; for other purposes they may yet remain viable and reliable information. The data may be accepted as is for some purposes, but rejected or accommodated for others.

Outlier problems must consider many interlocking relationships, as indicated by Barnett and Lewis (1978) and cannot be readily subdivided into any tidy topics. According to Barnett and Lewis (op. cit.), "A distinction must be made

1. between deterministic and statistical causes of outliers,
2. between univariate and multivariate data sets,
3. between different specific probability models,
4. between different forms of statistical analysis in which outliers are encountered,

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA, OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

5. between single or multiple outliers (including outlying samples and sets of samples), and

6. most fundamentally between the different aims and purposes we may have in studying outliers."

The gathering together of data measurements into groups which appear to describe and define objects of interest does provide a means upon which to base a decision, pro or con, or to make no decision at all. This "gathering" can be called gathering, typing, grouping, clustering, hierarchical development, patterning, condensation, coalescence, similarities, pattern recognition, dividing, clumping, typing, agglomeration, linkage, cohesions, analysis, and many other terms. Dissimilarities, discordancies, or differences rather than similarities seem to be the philosophic base for group determination. As Wolfe (1970) indicates, the many names used may originate in the many diverse disciplines. In addition, the naming is a function of the national language, the national ethos, the discipline language, as well as the concept maker and/or use of the results. The provincialism of the disciplines and the driving (forcing) function of doctoral programs, lead towards exploring one small bit of a problem, its exploitation, and the development of rival sets of words or phrases. Thus, development in one discipline may parallel that in another discipline, yet essentially be not recognized because of the word and language differences. For example, in mathematics, to the dismay of some undergraduates and graduates, the following are all synonymous: eigenvalues - eigenvectors, latent roots - latent vectors, characteristic roots - characteristic vectors, direction cosines, and roots, and more. Some of the above may also have other meanings.

Careful consideration of the words selected will provide an idea of concept of the authors and of the suggested metrics used or to be used. The concepts may influence or even stipulate the procedural steps. "Once a 'gather' has been made, some way to assign a measurement to one or more 'gathers' is usually sought." As mentioned previously, many terms covering many techniques are used. Some additional techniques are: aggregation; transcendental classification; dendritic, tree branch, root analysis; discriminant function analysis, transcendental function analysis, linear, quadratic, polynomial analysis; factor analysis; numerical taxonomy; and principal component analysis.

In nearly every one of the above techniques, computer programs permit the option for the use of "a priori" information. The use of such information is implied by the terms "learned" and "unlearned," "taught" and "untaught," or any other equivalent set of terms. "A priori" information should be used wherever possible, for it enhances the program operation from the decision criteria points of view as well as from the viewpoint of time and cost. Some procedures permit options in the assumptions and in the metrics that may be selected and used. The data themselves may suggest seemingly natural categories for the ensemble. The addition of data may change these seemingly natural categories. The word "suggest" is necessary because a different beginning from a perceptual viewpoint or the actual manipulation of data may lead to slightly different categories.

In summary, there are many ways to arrive at the grouping of data. Crutcher (1960) used ordered variances of the individual climatological and precipitation elements with various ratio tests, gap tests, and other tests before proceeding to the k-means testing (Bartlett 1937). Other investigators work with distances or metrics proceeding through the k-means tests and the establishment of the various configurations. With some of the techniques (Crutcher 1960 and McRae 1973), an observation (measurement) may enter and leave a cluster during the grouping process and do so repeatedly. With most hierarchical aggregation routines, however, once an observation enters a group it does not leave the group.

The basic assumption made by many investigators (Crutcher 1960 and Wolfe 1967a) is that the underlying distribution is multivariate-multimodal-normal or approximately normal. The techniques are sufficiently robust so that some departure from normality does not vitiate or change the decisions. On the other hand, it is the metric used that determines the groups with no reference to a distribution. In fact, the underlying statistical distributions in most programs are not known or may not even be considered. In general, any of the above techniques that are multivariate in nature may be used to look at the inherent structure of the data. From here on in this report, the words "cluster" or "clustering" will be used to denote the establishment of groups of data that optimize some selected criterion. Other words may be used when they are especially pertinent to the discussion.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

1.3 PREVIOUS WORK ON CLUSTERING

Mixed distributions have always presented confusing aspects to those, who for decision purposes, want to have some simple modes to enable them to make decisions. Grouping and classification problems have been worked on over many thousands of years. Only recently have statistical workers attacked the problem in specific ways. Impetus was given by Barnard (1936) and particularly by Fisher (1936), when discrimination techniques developed the important step of assignment or classification of data. Assignment of data to groups or clusters was made after appropriate statistical characteristics of the clusters had been determined.

A metric is defined to be some measure of distance between or among points or clusters of points. From the viewpoint of consistency and "a posteriori" conceptualization of relationships among the variates, the use of the same metric is advisable to group and to classify data. It is most difficult, if not impossible, for most researchers to visualize clustering in one metric and understand classification made by another person with a different metric in an "a priori" sense. Some people will accept a classification of objects without question while others will require an explanation of the metric used to establish the clusters.

The avidity of research workers to use a new technique to help them in their research is a measure of its implied or real importance. According to Fix and Hodges (1951), by 1950 less than 15 years after the introduction of the linear discriminant function (LDF) analysis techniques by Fisher (op. cit.), the literature references to LDF analysis numbered more than 200. Duran and Odell (1974) provide over 400 references to cluster analysis. An all inclusive reference or bibliographic set would be possible in an extensive and continuing study. The references for clustering provided in this report are minimum in number. No slight is intended to any researcher who has available pertinent papers which have not been mentioned.

Rapid wave-like advancements in scientific as well as in other fields such as commerce, agriculture, and industry is not unusual. They seem to emanate from some center of dissemination, such as a teaching school, an artist, a book, a university, a seminar, or just a group discussion. Oftentimes, the responsible person or group is at first unknown and may remain so. Witness

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

the use of pseudonyms such as "Student." Today, the distribution described by Gossett (1925) is known as "Students t-distribution" or as the t-distribution. What is essential is the development of an infrastructure.

Credit due for the theory of evolution is commonly attributed to Darwin. To give credit when credit is due is not always straightforward. The originator of a theory or technique is often passed over for whoever further develops or applies theory or technique. Those who apply the developments and let others know, most often receive the credit. Sometimes, a technique is developed in response to the impetus or the existence of a problem requiring a solution. The technique may be completely different or differ in only a few aspects yet still deserve treatment and discussion. Refinements on these techniques result in secondary wave advances. Thus, the early work of Pearson (1894, 1901) and Student (1925), pseudonym of Gosset, was followed by work culminating in the 1930's by Hotelling (1936), Mahalanobis (1936), and Fisher (1936). Apparently stemming from the same needs and with the impetus of Pearson's and Student's work, Hotelling (1954) showed the relationship among Hotelling's "T^2," Mahalanobis' "D^2," and Fisher's "LDF."

1.4 UNRESOLVED PROBLEMS IN CLUSTER ANALYSIS

1.4.1 Order of Clustering, Factor Analysis, and Regression Analysis

Everitt (1979) discusses unsolved problems in cluster analysis. He cites the work of Cormack (1971) and Everitt (1974) for review on cluster analysis. He makes the same point as Crutcher and Joiner (1977) with respect to linear regression, namely that cluster analysis should precede factor or regression analysis. Factor analysis and regression analysis are often used to provide input to cluster analysis, so that "a priori" information is provided for cluster analysis. In any case, reliable factor analysis and/or regression analysis should always follow for each cluster in any stage of groupings. Everitt further discusses problems associated with discriminant function analysis, analysis of variance, and graphic techniques.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

1.4.2 Scaling

A problem (Everitt 1979) encountered is the usual one of scaling which is not easy to solve. How should the variates be scaled? Many different ways to scale are possible for an infinite number of metrics exist or can be developed.

1.4.3 Stability

How should clusters be tested for reality, stability, and validity? Where there may be a question, should alternative clustering techniques be compared? A procedure permitting comparison of alternative clustering techniques would possibly permit the determination of better metrics as well as better clustering models. However, even from field to field such procedures may well present seemingly different answers and require additional work and thought.

Does the structure change when additional variates are used? Does the structure change when more data (variables) are added? Do the same group characteristics emerge if a new set of measurements are made? Usually, data sets are so limited in scope that these questions cannot be answered, yet investigators are content to simply report what they find and let it go at that. They may be presenting only one small facet of a much larger structure. As Everitt (1979) indicates, the report type depends upon whether the researcher is interested in dissection or in classification. An important point also made by Everitt (op. cit.) is "that many users of clustering are not clear in whether they are interested in dissection or classification and are not helped (in some cases apparently not bothered) by the lack of a satisfactory statistical test for distinct clusters."

1.4.4 Significance Assessment

How should cluster significance be assessed? In a subjective way, the significance can be assessed from the practical viewpoint. Grubbs (1950, 1969) discusses the outlier problem in this context. Outliers, singly or in groups, are clusters. Does the clustering do a reasonable and logical job? In a theoretical sense, appropriate models can be used for comparison, even though exact probabilities may never be known.

9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

1.4.5 Linkage

Single linkage ideas, pro and con, are discussed by Everitt (1979). The reader is referred to the literature for other ideas, concepts, and definitions not provided in this paper.

Investigators may or may not be concerned with the number of groups or clusters. However, even an entire tree structure in hierarchical grouping is strongly influenced by initial or starting points. In these and other hierarchical clustering schemes, it is the criterion of distance that appears to be the primary consideration. For example, when considering star clusters, is the metric chosen against the viewing plane of space or from the view of outer space towards the earth a straight line, great circle, or some other underlying conic? Is another metric better? The linkages and the clusters will be different depending on the metric chosen.

Everitt (1979) states with respect to linkages: "In general, the result of studies indicated that

1. no single method is best in every situation,
2. the mathematical single linkage is, in most cases, the least successful for the data used and with respect to other metrics, and
3. group average clustering and a method due to Ward (1963) do fairly well overall."

1.4.6 Covariance Matrices

The use of the minimization of the trace (W), the Within group covariance matrix, has several drawbacks, including that the method is transformation dependent and the clusters obtained are constrained to being hyperspheric even though they may be hyperspheroidal, i.e., clusters are forced to be circles or spheres although they may in reality be ellipses or ellipsoids. The minimization of the determinant (W), det (W), does not restrict clusters to being hyperspherical but does constrain the clusters to have the same shape. This constraint occurs in the Wolfe routines when the assumption of equal covariance matrices option is selected. The writer believes that not every set can be expected to exhibit the same shape and orientation in every subset or cluster as is obtained when the covariance matrices are identical.

Other techniques used the ratio B / W where "B" is the Between groups covariance matrix and $T = B + W$, where, respectively, these are the "Total," "Between" and "Within" sums of squares and cross-products. They are called pxp matrices. Some techniques use the relationship known as Wilks' Lambda, T / W.

1.4.7 Number of Groups

Another problem is the number of groups constituting a set of data. From a conceptual point of view, in the writer's opinion, the number of groups depends upon the underlying mixed distribution which may or may not be known and must be estimated. In light of the available data, of the background experience of those involved, and of the use to which the results are to be made, the metrics chosen will influence the number of groups obtained. It must be kept in mind that the greatest number of groups could be a cluster for every point. It is the purpose of clustering programs when exercising the principles of parsimony to reduce the "n" cluster concept to a lesser number of more meaningful clusters.

The number of groups may be 1) arbitrarily chosen as in the multivariate iterative k-cluster analysis, MIKCA McRae (1973) or 2) by decision criteria employed in Wolfe's NORMIX (1971) routines to determine clusters. In the writer's opinion, the establishment of an arbitrary number of clusters may well force formation of clusters that contain no salient information.

1.4.8 Efficacy of Wolfe's Technique

Binder (1978) using Binder's previous work as an authority and Everitt (1979) indicate that Wolfe's (1970) derived likelihood ratio test is not a valid test. Binder (op. cit.) indicates that the test does not arbitrarily have an asymptotic chi-square distribution and that the null distribution is often intractable. For assessing the null hypothesis that (k') groups are not significantly different from (k) groups, where (k'-k) is equal to or greater than 1, Wolfe's test was derived after examination of some Monte Carlo simulations. The writer chooses to accept Wolfe's test even though the probability levels may not necessarily be those implied by the user's selection of a probability level for rejection of the null hypothesis. Until more extensive

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Monte Carlo simulations are made and/or other more satisfactory tests are developed, there will be arguments as to the definitive shortcoming of all tests developed or suggested. Lee (1979) has developed a technique that should be examined to determine whether it does a better job than Wolfe's procedures. However, Lee's technique is restricted to the case of equal covariance matrices.

As clusters mean many things to many people, and the concepts vary as to the metrics to be used to describe the cluster structure, there may be no simple universally acceptable solution.

1.5 ELECTRONIC COMPUTER PROGRAMS AVAILABLE TO EPA-NCC AT RTP

In the fields of climatology and weather, Crutcher (1960) and Miller (1962), respectively, utilize grouping and clustering procedures in order to later discriminate and classify data. Though some early computer capability was available to Crutcher (1960), the magnitude of the problem required processing the data in separate steps. Increasing power and capability, and the availability of electronic computers allowed Miller (op. cit.) to also exploit the techniques of discrimination and classification (or prediction). The year a paper is presented lags the actual work by three to five years. As indicated previously, Wolfe (1970, 1971b) reveals progress of an individual research worker over a decade. Crutcher and Joiner (1977, 1978) and Crutcher (1979) adapted and used the NORMIX program of Wolfe (1971b) to study sets of climate and weather data. This program, which has been further modified in this project, will be made a unit of the library programs available to the U.S. Environmental Protection Agency National Computer Center (EPA-NCC) at the Research Triangle Park location.

Some other cluster programs do exist in the electronic computer statistical package programs in the EPA-NCC at RTP. Subsequent work will compare some of these programs both amongst themselves and with the Wolfe-Crutcher-Joiner-Graves-Stephenson modified NORMIX. The original Wolfe (1967a) NORMIX program was modified by Crutcher and Joiner in adapting it to the then RCA and now UNIVAC Spectra 70/45. This computer is no longer available at the National Climatic Center (NCC), Asheville, North Carolina. The original version is for the IBM 360 systems. The present version is for the UNIVAC

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

1110. Many options prior and post to the central NORMIX program have been added.

With the exception of CLUSTAN the statistical packages have embedded clustering programs. Only the MIKCA (McRae, 1973) program has been examined briefly.

The programs listed below are available at the NCC library to test and compare efficiency with the Wolfe NORMIX routine, although considerable manipulation will probably be necessary. NORMIX considerably modified, has been tested as a result of this present work, and is available for use by research investigators. NORMIX, without the adaptions and modification, is available at the Triangle University Computing Center (TUCC), Research Triangle Park, North Carolina. The computers are the UNIVAC 1110 and the IBM 360-165, respectively, at EPA-NCC and TUCC.

	Acronym	Name	Available	Tested
1.	BMDP	Biomedical Data Package	yes	no
2.	CLUSTAN	CLUSTER ANALYSIS	no	no
3.	IMSL	International Mathematical and Statistical Laboratories O-Cluster Linkage, O-Clink O-D-NORM, O-D-FISH	yes	no
4.	MIKCA	Multivariate Iterative K-Cluster Analysis	yes	no
5.	SAS CLUSTER	Statistical Analysis System CLUSTER	yes	no
6.	SPSS BREAKDOWN	Statistical Package for the Social Sciences BREAKDOWN	yes	no

The brief descriptions below are in general taken from those in the statistical data packages themselves.

1.5.1 BMDP

1. The acronym is derived from Biomedical Data Package. The clustering program is described in chapter 17, pages 621-652, as P1M, P2M, and P3M, Cluster Analysis of Variables, of Cases and Blocks, respectively. They are discussed in Hartigan (1975).

2. P1M forms clusters of variates based on a measure of association or similarity between the variables (such as correlation) or on a measure of distance separating the variables. The clusters are formed according to a specified linkage (or amalgamation) rule. Initially, each variable or vector of variables is considered to be a cluster that contains only that variable (vector). At each step the two closest or most similar clusters are joined to form a new cluster. Eventually all data are linked together to form one cluster. A similar procedure essentially is followed by all the hierarchical clustering algorithms for MIKCA. A summary table also is printed.

Diagrams such as the trees or mapping output allow the user to determine the last point (datum) or small cluster to join the main body of data.

3. P2M uses one of four distance metrics to create clusters in the same fashion as in P1M. However, this program is for cases, not variables.

4. P3M combines the capabilities of P1M and P2M.

5. In P1M, the measure of similarity can be the correlation between variables, the absolute value of the correlation, or a measure of the distance between two variables such as the arccosine. The criterion used to combine the variables into clusters, or clusters into larger clusters, can be based on 1) the average measure of similarity (average linkage) between the variables or clusters, 2) the maximum similarity (single linkage), or 3) the minimum similarity (complete linkage).

This clustering program algorithm may possibly do as good a job as Ward's (1963) algorithm embedded in the NORMIX program. No test has been made as yet.

1.5.2 CLUSTAN

1. The acronym CLUSTAN is derived from the two words CLUSTER ANalysis. The University of Research Councils series, Report No. 47, January, 1978, Users Manual, Third Edition, pp. 1-175, describes this program (Wishart 1978).

2. This statistical data package is not available at EPA-NCC. The programs are linked with BMDP and SPSS.

3. This program is an extensive package of statistical computing algorithms. It would be a valuable adjunct for any computer library with the capability of that at EPA-NCC.

4. Its best clustering algorithm is that of Ward (1963), which is also contained in NORMIX. Therefore, its output in that sense will not be any better than that of NORMIX. However, it does have many other techniques that may appeal to research workers. Comparisons of the output using different metrics can be made. The program is not discussed here in any detail.

1.5.3 IMSL

1. IMSL is the acronym of the statistical data package of the International Mathematical and Statistical Laboratories IMSL LIB-0006, November 1977, Edition 6.

Three Chapters O, S, and U contain algorithms identified further with the acronyms, OCLINK, SSSCAN, and USTREE. OCLINK refers to linkage or complete linkage with hierarchical clustering analysis given a similarity matrix. SSSCAN refers to population mean and total single stage cluster sampling with continuous data. USTREE refers to printing a binary tree that will represent the output of OCLINK if the output of OCLINK is used as input.

2. The subroutine OCLINK provides hierarchical clustering capability via either single linkage or complete linkage algorithms. A similarity matrix, that may be distance-like or correlation-like, is input to OCLINK. The user may choose the desired measure of similarity for entry to OCLINK by processing the data through prior routines. Subroutine USTREE may be called to print the tree.

3. The hierarchical clustering algorithm is simple. With a given number of points in p-dimensional space, where each point is a cluster, a new cluster is formed by combining the two closest points. The similarity level is recorded. The similarity matrix is updated by calculating the similarities (distances) between the new cluster and all the other clusters (points). The process is repeated until all points have merged into one cluster.

4. In data validation, it will be the last singleton point or last small sized cluster to merge with others that has the highest potential of being classed as an outlier. It then can be accommodated, modified, accepted, rejected or corrected.

5. These groups and their characteristics can then be processed through discriminant analysis to provide classification or prediction. These analysis routines are ODNORM and ODFISH. ODNORM is multivariate linear discriminant analysis among the several known groups produced by OCLINK or by other clustering algorithms. ODFISH is the linear discriminant function (LDF) analysis of Fisher to reduce the number of variables. These variables are different from the input variables and are combinations of the input variables. The output of the techniques from these programs have not been investigated here by the writer. They should be investigated. However, in order to make the output comparable to NORMIX, better estimates of the statistical parameters may have to be obtained than the clustering procedures provide. The clustering output is not believed to be any better than and perhaps not as good as those available in any routine that can utilize Ward's (1963) hierarchical clustering scheme.

1.5.4 MIKCA

1. The acronym MIKCA stands for Multivariate Iterative K-Cluster Analysis. McRae (1973) published MIKCA; M.J. Symons revised MIKCA in 1973 and is again revising it.

2. The revised version of MIKCA is not yet available at the EPA-NCC library. The older version is available on the TUCC computer.

3. The number of clusters must be specified. Nine criteria which may be optimized are indicated. Criteria are measures of the within groups sums of squares and cross products matrix W. Pooling is effected. Only two of the following criteria appear to be working well at the present time and are indicated by the word "working"

-Trace W
-Determinant W or det W (working)
-Largest root of W (Inverse)
-Trace W (Inverse) times B
-Bayesian Modification of (2): Locally uniform (working)
-Bayesian Modification of Unequal Dispersion
-Maximum Likelihood of the Determinant W
-Maximum Likelihood for Unequal Dispersion Matrices
-Maximum Likelihood for Standardization of Parameters

The data may or may not be standardized.

4. The distance parameters are Euclidean, scaled Euclidean, and Mahalanobis. Distances are measured from cluster to cluster center.

5. The procedure may utilize random numbers to initiate clustering. They must be extracted from the next to final sets before the final sets can be prepared.

6. The clustering is not strictly hierarchical for data vectors can move from cluster to cluster. By the criteria and metrics used, as the clusters coalesce, a data vector which appeared to belong to a specific cluster may no longer appear to belong and moves to a more "friendly" neighborhood, i.e., the nearer cluster to which it best appears to belong, or it remains a singleton cluster.

7. Fairbairn, Nelson, and Symons (1979) discuss this program and use it to describe the distribution of data in clusters from the Los Angeles Catalyst Study (LACS) by EPA (1977). In addition, they compare the output with the SAS clustering algorithm output. The NORMIX technique is briefly discussed and recommended for future exploration and use. Dendritic diagrams are shown. Mapping output, similar to that in the NORMIX output, is illustrated and discussed. The study of LACS data by use of the MIKCA program involved the clustering of the following pollutant and meteorological data: suspended particles, lead, carbon monoxide, windspeed, wind direction, temperature, and the additional variable, traffic count. Five outliers and their characteristics are discussed. The reader is referred to their report for more details.

1.5.5 SAS

1. The main acronym is SAS and the secondary acronym is CLUSTER. SAS is a Statistical Analysis System computer software package. It contains many different computing routines among which is CLUSTER.

2. This program is available in the EPA-NCC library at RTP. The computer is the UNIVAC 1110.

3. CLUSTER computes its own distance matrix. The metric is Euclidean. The CLUSTER procedure performs a hierarchical cluster analysis. One cluster for each observation is established. The two closest point clusters are combined into a cluster. Then the closest existent two clusters, point cluster, or a point cluster and the just formed two-point cluster are joined.

This procedure is continued until only one cluster, the total data set is attained.

4. All of the clustering techniques discussed in this present paper are of the agglomerative or aggregative type and not of the divisive type.

Optionally, the distances may be standardized by dividing each element by the average distance from the vector of variance means. This transformation is especially good for gamma distributed variates.

The distance between two clusters is defined to be the maximum distance between an observation in one cluster and an observation in the other cluster. This definition is just one way to compute distances. Alternatively, they could be minimum distances, distances between centroids, or such distances divided by some measure of dispersion such as the standard deviation.

5. CLUSTER should not be applied to a data base set of more than 250 observations. CLUSTER will print a summary of the cluster analysis performed. For each step, it prints the number of cluster, the maximum distance between two observations in a cluster, the number of other distances within a cluster, all of which are less than the maximum diameter. The number of such distances is equal to (n(n-1)/2), where the total number of distances is less than the maximum diameter and the ratio of the last two quantiles. As an option, observations in a cluster are joined by a line. If a specific number of clusters is given, the procedure will print the minimum, average, and maximum distances within and between clusters, a listing of the observations within each cluster, and the means of the variables within each cluster.

6. Barr, Goodnight, Sale, and Hellwig (1976) discuss the SAS data package.

1.5.6 SPSS

1. The main acronym is SPSS and the secondary acronym is BREAKDOWN. SPSS, the Statistical Package for the Social Sciences, contains many different computing routines.

2. This program is available in the EPA-NCC library at RTP. The computer is the UNIVAC 1110.

3. This program can use integers or continuous variates.

4. BREAKDOWN does not give a clustering in the same sense as NORMIX. The clusters are often arbitrarily assigned. However, this arrangement may be applicable for certain aspects of air pollution problems, instrumentation, data validation, etc. As this program has not been tested nothing can be written here to tell what this program does. It should be examined and tested to determine its usefulness and efficacy as compared to NORMIX and other clustering routines.

6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

2.0 DATA VALIDATION

2.1 ROGUE OR OUTLIER TESTS

There may be errors of omission or of commission when an observation yields a datum in one or more variates. In any data collective, until all data are available, if this is indeed possible, there may be a datum which does not appear to belong to the postulated or hypothesized model that appears more and more to be the most likely. Even after all data are available that are to be used, such outliers may still be present.

Uncertainties of parameter estimates lead to model (distribution) uncertainty. The imprecision generally decreases slowly with an increasing number of data. The impression of a datum or data subset of "belonging" or "not belonging" to a particular distribution will vary, depending on the distribution or the assumed distribution. This impression also depends upon the basic experience or whoever is collecting, working, or viewing the data.

Discussions and techniques (equations or algorithms to assess the extent of discordancy of data within many distributions) are presented by Barnett and Lewis (1979) and Green and Carroll (1976). They involve singleton or multiple outlying data subsets at either or both ends of univariate distributions. Multivariate outliers are also discussed.

The word "rogue" is used in this section's heading to better describe a datum that simply does not fit into an ordered multiple-dimensioned and well-structured field. An example of a rogue in such a field is the existence of an apparently discordant datum in a two or three dimensional field representing an electromagnetic field, a pressure field, a gravitational force field, or any similar type of field. The word "outlier" does not seem to be quite descriptive enough in this case. A specific example of a "rogue" can be found in an isobaric field where, on a latitude-longitude projection plane, surface air pressures are plotted on a weather map. In that plane, a smooth analysis would be attained except for the effect of one point, say 1010 mbs in a field ranging from 990 to 1000 mbs. If the pressure data were ordered without reference to the plane, the word "outlier" would suffice. For purposes of

BOTTOM OF IMAGE AREA, OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

this report, the word "outlier" will be used, although in the case cited above, in the viewing plane, it would be a discordant "inlier."

In many instances, the need for data set integrity is the driving reason for the study of outliers. Usually the initial work is an editing process to identify those data which are discordant in some fashion or another. As a first simple check, tabular arrays that permit screening for data having values which are physically impossible, etc., are generally used. Vast data banks and powerful computer arrays permit such editing which is in no sense quality control or assurance. These tabular arrays are simply gross checks, after the fact, of the gross quality of the process involved. For some of these techniques, see Crutcher (1979) and Klint (1979) as well as the total report, EPA Data Validation Conference Proceedings (1979). Individual totally unrealistic values are detected. The multivariate techniques do permit a check on more than one element at a time. Seeming outliers in the one dimensional sense may not be outliers in the multidimensional sense. Furthermore, outliers in the multivariate sense may not be outliers in any of the marginal single element distributions.

While the following discussion is simplistic, its application may be difficult in making penultimate decisions. The word penultimate is used rather than ultimate as decisions generally given in probabilistic terms are subject to challenge and therefore possible modification. Often models are used to effect decisions. There is a "truism" given in an infinite data set: any and every assumed model tested will be rejected for then the distribution is explicit, known with certainty and no approximating model will serve. Another "truism" is that the chance of a model exactly fitting or describing a random set (or vice-versa) is small indeed.

Any statistical test (or even a test made without benefit of statistics) has a null hypothesis and one or more alternative hypotheses. A decision is made that a datum is accordant, discordant, or no decision is made. If the null hypothesis allows for no outlier, no discordancy, an alternative hypothesis may suggest a way in which the model may be modified to use or explain the datum.

If we suggest that an observation is an outlier, we have some framework or model in mind, although we may not have specified the model. A null hypothesis is a statement or specification of the initial basic probability

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
9-1/8"
3/8"
EPA-287 (Cin.)
(4-76)

model. We may not need to be specific, or we can be so specific as to state that the data constitute a random sample from a distribution such as a gamma distribution. Even more specifically, an exponential, chi-square, Poisson, or normal distribution may be proposed for they are all approximated by or equivalent to special cases of the gamma distribution as well as the beta distribution. Alternative hypotheses concern deterministic, mixed, slippage, exchangeable, and other alternatives.

Given observations or data, $x_i (i=1,2,3...n)$, they could be single element (univariate) or vector (multivariate) data. This discussion is applicable to the multivariate problem or any one of its marginal elemental distributions. In essence, the distance, however defined between singleton x_i or subsets of x_i and the remaining general x_i configuration, is the factor used to establish whether the datum or data subset may be thought of as outlying. One procedure is to obtain a ratio of such a distance to some dispersion characteristic. This ratio provides a dimensionless yet meaningful number; ratio of dispersions is also useful. The usual "F" test may be involved. The point here is that any difference, i.e., distance, whether in the numerator or denominator, is some measure of the dispersion.

Suppose for instance, that simple Euclidean distance is used in a univariate context. The differences that could be computed for each data set is a triangular matrix of differences of a size of $(n(n-1)/2)$ terms. Any one of these sets could appear, as well as any other measure of the dispersion such as the population, the sample range, or sample standard deviation. The options are many. If subsets are used, the problem becomes more complex. Options exist in both numerator and denominator; therefore the ratio may increase in both the combinatorial and permutation sense. Many investigators have examined selected cases.

In some cases, in the examination of a singleton univariate potential outlier, such as in the case of ordered x_i, from low to high, the ratio $(x_n-\mu)/\sigma$ may be used, where μ and σ are the population mean and standard deviation, respectively. Where the mean and variance are both unknown and must be estimated, the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are equal, respectively, to $\bar{x}$ and $\Sigma (x_i - \bar{x})^2/n$, where i ranges from 1 to n.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Each test requires its own set of tables or algorithms for implementation of decisions as to whether the x_n is an outlier, what alternative hypothesis to examine, and then whether to reject (eliminate), correct, or modify the model, to accommodate the outlier or to accept the outlier as a separate cluster of viable and important information.

Dixon (1950) provides an excellent analysis of extreme values. He states (page 489), "The performance of the various criteria for discovery of one or more contaminators will be measured with reference to contaminators of the following two types entering into samples of observations from a normal population with mean μ and variance σ^2, $N(u, \sigma^2)$

A. One or more observations from $N(\mu + \lambda\sigma, \sigma^2)$

B. One or more observations from $N(\mu, \lambda^2\sigma^2)$."

Here, N implies the normal distribution and lambda (λ) is scalar, i.e., it has magnitude only. Inclusions of these criteria in either or both A and B will produce mixtures where they are either detected or at least their presence indicated by many tests designed to detect one or more of them. Sometimes they are referred to as shifts or slippages in the mean and/or the variance. The first (A) is a "location" error, while the second (B) is a "scalar" error, according to Dixon (1950).

Techniques based on the early work of Chauvenet (1863) and Wright (1884) did not always provide satisfactory results. Thompson (1935) indicated perhaps the first basic theoretical approach to the problem. He provides a table to use with his techniques. Grubbs (1950), Ferguson (1961), and Barnett and Lewis (1978) discuss these problems in detail. Grubbs (1950) also includes tables in his paper.

Thompson (1935) defines $\tau_i = (x_i - \bar{x})/s$ where $\bar{x}$ and s denote the sample arithmetic mean and the sample standard deviation, respectively. Thompson (op. cit.) indicates the relationship of τ_i to Student's distribution.

Pearson and Chandra Sekar (1936) show that the distribution of τ is a symmetrical limited range distribution of Pearson's Type II and further show that $t = \tau (N-2)/(N-1-\tau^2)^{\frac{1}{2}}$ by Thompson (op. cit.) follows Student's t distribution with (N-2) degrees of freedom. Appropriate tables are found in many texts. Some computer programs permit direct computation. These two papers are extremely interesting and illuminating.

Harter (1974-1976) provides an excellent reference survey of investigative work. Though many have studied the problem of outlying observations, no one has ever devised a "trouble-free" criterion for rejection of outlying or extreme observations.

In multivariate data sets, outliers are not so easy to spot by visual inspection. In fact, an outlier may be internal (an inlier) on each vector element but existent as an extreme in the total vector. On the other hand, an outlier in one or more vector elements may well be within bounds in the multivariate sense.

Those familiar with only the concept of the symmetrical normal distribution might well consider a subset consisting of one or more observations to be outlying, while a person familiar with asymmetrical models might not consider the data to be outlying. Thus, heavy disproportionate rainfall over short time periods is not as unexpected as over long time periods when the characteristics of each time period is considered. The underlying distributions of rainfall is usually described by the general gamma distribution, which is a subdistribution in itself, of the more general beta distribution. Both approximate normal distribution under special conditions.

An "outlier" datum, subset, or cluster may be valid or erroneous. Such an outlier may be accepted as being valid or set aside for certain purposes such as study on the outlying set or on the remainder of the data set. Barnett and Lewis (1979) state the problems facing the investigator when outliers are detected. They provide four courses of action:

1. reject or correct the discordant outliers,
2. modify the model to incorporate the outliers in a nondiscordant fashion,
3. refine the analysis procedures for the entire data set so as to accommodate the outliers, and

9-1/8"
BEGIN LAST LINE OF TEXT
3/8"
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

4. concentrate attention on the discordant outliers as a welcome identification of unsuspected factors of practical importance.

Key possible actions then are rejection or correction, modification, accommodation or acceptance, and the establishment of new kernels of information.

The discussion of the tests above may lead one to believe that the solution is simple and easy, but it is not. Most clustering programs, through some hierarchical scheme, create the clusters but do not provide any further information. Other programs go further and attempt to provide some estimate of the statistical parameters of the clusters. "NORMIX," for example, uses maximum likelihood estimation procedures.

Multivariate as well as univariate clustering or grouping techniques offer potential in data validation, data quality assurance, data quality control, as well as in classification and prediction.

The prime objectives of this report, as discussed above, are:

1. the provision of an electronic data computer program capable of clustering multivariate data, and

2. the detection and characterization of clusters which will permit data validation, and classification and/or prediction.

They are all associated with quality assurance of information gathering, hopefully, through quality controlled systems. These same techniques also can be used to monitor the systems producing the information.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

3.0 NORMAP-NORMIX

Wolfe (1965, 1967a, 1967b, 1968, 1970, 1971a, 1971b, 1971c) clearly shows more than a decade of work devoted to the development of techniques for the separation of heterogeneous data sets into clusters of more or less homogeneous characteristics. Two electronic data computer programs evolved during that time: NORMAP and NORMIX are the acronyms, respectively, for NORMAL MAPPING, and NORMAL MIXTURES. The first utilizes equal covariance matrix assumptions. Linear discriminant functions are developed and cluster identifier plots are provided in a field of one discriminant function versus another. For each cluster there is a linear discriminant function that is used for classification. The second utilizes the assumption of unequal covariance matrices. No discriminant functions are computed. If computed, they would be quadratic rather than linear (Crutcher 1960). Rather than to solve for the roots of the very complicated structural roots and to make decisions which involve the concept of unions and intersections, distances to the density surfaces are computed. Probability assignments are made on the basis of the relative distance magnitudes; no plots are made in the manner done with the linear discriminant functions. Extraction of distance information from the central computing processing could be done but would involve careful modification and considerable recalibration to assure that no unexpected changes have been introduced.

The program is now a combined program with a choice of one or the other of the two options above. NORMIX is now the acronym of the combined programs. Crutcher and Joiner (1977) modified the Wolfe (1967a) program for the RCA Spectra 70/45, now the Univac 70/45, at the NOAA-NCC library center in Asheville, North Carolina. That program in adapted machine language is no longer viable due to the replacement of the computer. For the present project, the program that duplicates the 1974 version of the Wolfe program has been extensively modified for the UNIVAC 1110 by the addition of many options and the use of double precision throughout the program. The options are grouped into 1) a data preparatory program prior to entry into the NORMIX central core computing section, and 2) an extensive set of output options.

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Some options are calls to statistical data packages. Other options developed for this program include output plotting programs. Secondary data files for the estimates of statistical parameters were established. Appropriate data are then extracted and processed to permit the plotting of probability distribution ellipses and of data profiles against a standardizing 3-sigma background. Other plots can be easily programmed and plotted with the data from these secondary bases as input information. A secondary data file of cluster assignments collated with the original data or transformed data would allow the computation of statistical parameter estimates without the necessity of developing a back-transformation for each transformation used. However, back-transformations are optionally provided.

Wolfe's publications include listings of the original program. Anyone desiring the 1974 version of the program should correspond directly with Wolfe at the address given in 1967 reference. Those programs are for the IBM series of electronic computers, IBM 360. The modified program, produced during this project for use on the UNIVAC 1110, will be documented in a report to follow. The detailed documentation of the modified NORMAP-NORMIX program including the various subprograms are published in Volume II of this report series.

Fisher (1936) used Anderson's (1935) set of Iris data of the Gaspé Peninsula in his work on discriminant function analysis. This same data set has been used so many times it has become a classic. There are 150 observations, 50 each on three Iris species: I. setosa, I. versicolor, and I. virginica. There are four measurements on each plant flower: sepal length, sepal width, petal length, and petal width. Wolfe (1967) used these data first to calibrate the NORMIX model and in 1968 to calibrate the NORMAP model. Others have done likewise with their models, as have we in the present report. A question has been raised in the literature concerning the classification of a few of the 4-variate data vectors. Two of the sets are considered to be distinct but the third appears to be a mixture, i.e., composed of hybrids perhaps exhibiting polyploidy. Various investigators have used different metrics to cluster and to classify. Each investigator, including myself, has

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

examined the classifications obtained by other investigators to determine whether the same results have been obtained.

Table 1 illustrates some of the input information required by the program. The number of data, the number of initial k-clusters to be used, the proposed level of testing of the null hypothesis, the identification of the variates and the number of clusters to be tested are some of the required salient input information.

```
                         NORMIX PROGRAM WITH IRIS DATA

                            VARIABLE NAMES AND UNITS

                  1      SEPAL LENGTH     SEP LNG          CM
                  2      SEPAL WIDTH      SEP WTH          CM
                  3      PETAL LENGTH     PET LNG          CM
                  4      PETAL WIDTH      PET WTH          CM

     NVBLES = 4  NSAMPL = 150  DIFF COV MATRIX
     MIN CLUSTER SIZE = 5  HYPOTHESIS TEST = .0500
     INITIAL KMEANS = 150  TIME LIMIT = 2  ITER LIMIT = 100
     NO. OF TYPES = 1   2   3   4   5   6   0   0   0   0   0   0   0
     FORMAT FOR READING DATA =   (6E12.5,/,6E12.5)
```

Table 2 provides the four element observations obtained by Anderson (1935) on the three Iris data sets that were used by Fisher (1936). They are listed vertically on the left in one columnar set paired with two columnar classification sets on the right. This table represents the input data and their classification after processing. These results can be compared with the results of Fisher (1936) and Wolfe (1967a, 1968) and were obtained by use of the modified program now available in the EPA-NCC at RTP. Only those output parts pertinent to the understanding of the modified program will be presented here. For further details, the reader (user) is referred to the investigators' work referenced immediately above and to the forthcoming documentation.

Table 2. Fisher Iris data; input data and output classifications for 3 clusters, I. setosa, 1-50, I. versicolor, 51-100, and I. virginica, 101-150 with sepal length, sepal width, petal length, and petal width

```
NORMIX PROGRAM WITH IRIS DATA                                   PROBABILITIES OF TYPE MEMBERSHIP

     INPUT VARIABLES                                               UNEQUAL                       EQUAL
SEQ         1           2           3           4            COVARIANCE MATRICES        COVARIANCE MATRICES

                                                                1      2      3            1      2      3

@ADD,P IRIS2.
     1   5.1000      3.5000      1.4000       .2000         1 1.000  .000  .000         1 1.000  .000  .000
     2   4.9000      3.0000      1.4000       .2000         2 1.000  .000  .000         2 1.000  .000  .000
     3   4.7000      3.2000      1.3000       .2000         3 1.000  .000  .000         3 1.000  .000  .000
     4   4.6000      3.1000      1.5000       .2000         4 1.000  .000  .000         4 1.000  .000  .000
     5   5.0000      3.6000      1.4000       .2000         5 1.000  .000  .000         5 1.000  .000  .000
     6   5.4000      3.9000      1.7000       .4000         6 1.000  .000  .000         6 1.000  .000  .000
     7   4.6000      3.4000      1.4000       .3000         7 1.000  .000  .000         7 1.000  .000  .000
     8   5.0000      3.4000      1.5000       .2000         8 1.000  .000  .000         8 1.000  .000  .000
     9   4.4000      2.9000      1.4000       .2000         9 1.000  .000  .000         9 1.000  .000  .000
    10   4.9000      3.1000      1.5000       .1000        10 1.000  .000  .000        10 1.000  .000  .000
    11   5.4000      3.7000      1.5000       .2000        11 1.000  .000  .000        11 1.000  .000  .000
    12   4.8000      3.4000      1.6000       .2000        12 1.000  .000  .000        12 1.000  .000  .000
    13   4.8000      3.0000      1.4000       .1000        13 1.000  .000  .000        13 1.000  .000  .000
    14   4.3000      3.0000      1.1000       .1000        14 1.000  .000  .000        14 1.000  .000  .000
    15   5.8000      4.0000      1.2000       .2000        15 1.000  .000  .000        15 1.000  .000  .000
    16   5.7000      4.4000      1.5000       .4000        16 1.000  .000  .000        16 1.000  .000  .000
    17   5.4000      3.9000      1.3000       .4000        17 1.000  .000  .000        17 1.000  .000  .000
    18   5.1000      3.5000      1.4000       .3000        18 1.000  .000  .000        18 1.000  .000  .000
    19   5.7000      3.8000      1.7000       .3000        19 1.000  .000  .000        19 1.000  .000  .000
    20   5.1000      3.8000      1.5000       .3000        20 1.000  .000  .000        20 1.000  .000  .000
    21   5.4000      3.4000      1.7000       .2000        21 1.000  .000  .000        21 1.000  .000  .000
    22   5.1000      3.7000      1.5000       .4000        22 1.000  .000  .000        22 1.000  .000  .000
    23   4.6000      3.6000      1.0000       .2000        23 1.000  .000  .000        23 1.000  .000  .000
    24   5.1000      3.3000      1.7000       .5000        24 1.000  .000  .000        24 1.000  .000  .000
    25   4.8000      3.4000      1.9000       .2000        25 1.000  .000  .000        25 1.000  .000  .000
    26   5.0000      3.0000      1.6000       .2000        26 1.000  .000  .000        26 1.000  .000  .000
    27   5.0000      3.4000      1.6000       .4000        27 1.000  .000  .000        27 1.000  .000  .000
    28   5.2000      3.5000      1.5000       .2000        28 1.000  .000  .000        28 1.000  .000  .000
    29   5.2000      3.4000      1.4000       .2000        29 1.000  .000  .000        29 1.000  .000  .000
    30   4.7000      3.2000      1.6000       .2000        30 1.000  .000  .000        30 1.000  .000  .000
    31   4.8000      3.1000      1.6000       .2000        31 1.000  .000  .000        31 1.000  .000  .000
    32   5.4000      3.4000      1.5000       .4000        32 1.000  .000  .000        32 1.000  .000  .000
    33   5.2000      4.1000      1.5000       .1000        33 1.000  .000  .000        33 1.000  .000  .000
    34   5.5000      4.2000      1.4000       .2000        34 1.000  .000  .000        34 1.000  .000  .000
    35   4.9000      3.1000      1.5000       .2000        35 1.000  .000  .000        35 1.000  .000  .000
    36   5.0000      3.2000      1.2000       .2000        36 1.000  .000  .000        36 1.000  .000  .000
    37   5.5000      3.5000      1.3000       .2000        37 1.000  .000  .000        37 1.000  .000  .000
    38   4.9000      3.6000      1.4000       .1000        38 1.000  .000  .000        38 1.000  .000  .000
    39   4.4000      3.0000      1.3000       .2000        39 1.000  .000  .000        39 1.000  .000  .000
    40   5.1000      3.4000      1.5000       .2000        40 1.000  .000  .000        40 1.000  .000  .000
    41   5.0000      3.5000      1.3000       .3000        41 1.000  .000  .000        41 1.000  .000  .000
    42   4.5000      2.3000      1.3000       .3000        42 1.000  .000  .000        42 1.000  .000  .000
    43   4.4000      3.2000      1.3000       .2000        43 1.000  .000  .000        43 1.000  .000  .000
    44   5.0000      3.5000      1.6000       .6000        44 1.000  .000  .000        44 1.000  .000  .000
    45   5.1000      3.8000      1.9000       .4000        45 1.000  .000  .000        45 1.000  .000  .000
```

(continued)

Table 2 (continued)

					PROBABILITIES OF TYPE MEMBERSHIP							
					UNEQUAL COVARIANCE MATRICES				EQUAL COVARIANCE MATRICES			
	1	2	3	4		1	2	3		1	2	3
46	4.8000	3.0000	1.4000	.3000	46	1.000	.000	.000	46	1.000	.000	.000
47	5.1000	3.8000	1.6000	.2000	47	1.000	.000	.000	47	1.000	.000	.000
48	4.6000	3.2000	1.4000	.2000	48	1.000	.000	.000	48	1.000	.000	.000
49	5.3000	3.7000	1.5000	.2000	49	1.000	.000	.000	49	1.000	.000	.000
50	5.0000	3.3000	1.4000	.2000	50	1.000	.000	.000	50	1.000	.000	.000
51	7.0000	3.2000	4.7000	1.4000	51	.000	1.000	.000	51	.000	1.000	.000
52	6.4000	3.2000	4.5000	1.5000	52	.000	.999	.001	52	.000	.999	.001
53	6.9000	3.1000	4.9000	1.5000	53	.000	.994	.006	53	.000	.997	.003
54	5.5000	2.3000	4.0000	1.3000	54	.000	.930	.070	54	.000	1.000	.000
55	6.5000	2.8000	4.6000	1.5000	55	.000	.970	.030	55	.000	.997	.003
56	5.7000	2.8000	4.5000	1.3000	56	.000	.968	.032	56	.000	.998	.002
57	6.3000	3.3000	4.7000	1.6000	57	.000	.986	.014	57	.000	.979	.021
58	4.9000	2.4000	3.3000	1.0000	58	.000	1.000	.000	58	.000	1.000	.000
59	6.6000	2.9000	4.6000	1.3000	59	.000	.999	.001	59	.000	1.000	.000
60	5.2000	2.7000	3.9000	1.4000	60	.000	.966	.034	60	.000	.999	.001
61	5.0000	2.0000	3.5000	1.0000	61	.000	.998	.002	61	.000	1.000	.000
62	5.9000	3.0000	4.2000	1.5000	62	.000	.993	.007	62	.000	.999	.001
63	6.0000	2.2000	4.0000	1.0000	63	.000	.999	.001	63	.000	1.000	.000
64	6.1000	2.9000	4.7000	1.4000	64	.000	.967	.033	64	.000	.994	.006
65	5.6000	2.9000	3.6000	1.3000	65	.000	1.000	.000	65	.000	1.000	.000
66	6.7000	3.1000	4.4000	1.4000	66	.000	1.000	.000	66	.000	1.000	.000
67	5.6000	3.0000	4.5000	1.5000	67	.000	.924	.076	67	.000	.967	.033
68	5.8000	2.7000	4.1000	1.0000	68	.000	.996	.004	68	.000	1.000	.000
69	6.2000	2.2000	4.5000	1.5000	69	.000	.003	.997	69	.000	.979	.021
70	5.6000	2.5000	3.9000	1.1000	70	.000	.999	.001	70	.000	1.000	.000
71	5.9000	3.2000	4.8000	1.8000	71	.000	.053	.947	71	.000	.133	.867
72	6.1000	2.8000	4.0000	1.3000	72	.000	1.000	.000	72	.000	1.000	.000
73	6.3000	2.5000	4.9000	1.5000	73	.000	.042	.958	73	.000	.868	.132
74	6.1000	2.8000	4.7000	1.2000	74	.000	.912	.088	74	.000	1.000	.000
75	6.4000	2.9000	4.3000	1.3000	75	.000	1.000	.000	75	.000	1.000	.000
76	6.6000	3.0000	4.4000	1.4000	76	.000	1.000	.000	76	.000	1.000	.000
77	6.8000	2.8000	4.8000	1.4000	77	.000	.988	.012	77	.000	.999	.001
78	6.7000	3.0000	5.0000	1.7000	78	.000	.329	.671	78	.000	.704	.296
79	6.0000	2.9000	4.5000	1.5000	79	.000	.964	.036	79	.000	.991	.009
80	5.7000	2.6000	3.5000	1.0000	80	.000	1.000	.000	80	.000	1.000	.000
81	5.5000	2.4000	3.8000	1.1000	81	.000	.999	.001	81	.000	1.000	.000
82	5.5000	2.4000	3.7000	1.0000	82	.000	1.000	.000	82	.000	1.000	.000
83	5.8000	2.7000	3.9000	1.2000	83	.000	1.000	.000	83	.000	1.000	.000
84	6.0000	2.7000	5.1000	1.6000	84	.000	.007	.993	84	.000	.127	.873
85	5.4000	3.0000	4.5000	1.5000	85	.000	.848	.152	85	.000	.929	.071
86	6.0000	3.4000	4.5000	1.6000	86	.000	.987	.013	86	.000	.988	.012
87	6.7000	3.1000	4.7000	1.5000	87	.000	.997	.003	87	.000	.999	.001
88	6.3000	2.3000	4.4000	1.3000	88	.000	.918	.082	88	.000	1.000	.000
89	5.6000	3.0000	4.1000	1.3000	89	.000	.999	.001	89	.000	1.000	.000
90	5.5000	2.5000	4.0000	1.3000	90	.000	.990	.010	90	.000	1.000	.000
91	5.5000	2.6000	4.4000	1.2000	91	.000	.940	.060	91	.000	.999	.001
92	6.1000	3.0000	4.6000	1.4000	92	.000	.991	.009	92	.000	.998	.002
93	5.8000	2.6000	4.0000	1.2000	93	.000	.999	.001	93	.000	1.000	.000
94	5.0000	2.3000	3.3000	1.0000	94	.000	1.000	.000	94	.000	1.000	.000
95	5.6000	2.7000	4.2000	1.3000	95	.000	.994	.006	95	.000	1.000	.000
96	5.7000	3.0000	4.2000	1.2000	96	.000	.997	.003	96	.000	1.000	.000
97	5.7000	2.9000	4.2000	1.3000	97	.000	.998	.002	97	.000	1.000	.000
98	6.2000	2.9000	4.3000	1.3000	98	.000	.999	.001	98	.000	1.000	.000
99	5.1000	2.5000	3.0000	1.1000	99	.000	1.000	.000	99	.000	1.000	.000
100	5.7000	2.8000	4.1000	1.3000	100	.000	.999	.001	100	.000	1.000	.000
101	6.3000	3.3000	6.0000	2.5000	101	.000	.000	1.000	101	.000	.000	1.000
102	5.8000	2.7000	5.1000	1.9000	102	.000	.000	1.000	102	.000	.001	.999
103	7.1000	3.0000	5.9000	2.1000	103	.000	.000	1.000	103	.000	.000	1.000

(continued)

Table 2 (continued)

						PROBABILITIES OF TYPE MEMBERSHIP						
						UNEQUAL COVARIANCE MATRICES				EQUAL COVARIANCE MATRICES		
	1	2	3	4		1	2	3		1	2	3
104	6.3000	2.9000	5.6000	1.8000	104	.000	.000	1.000	104	.000	.001	.999
105	6.5000	3.0000	5.8000	2.2000	105	.000	.000	1.000	105	.000	.000	1.000
106	7.6000	3.0000	6.6000	2.1000	106	.000	.000	1.000	106	.000	.000	1.000
107	4.9000	2.5000	4.5000	1.7000	107	.000	.000	1.000	107	.000	.022	.978
108	7.3000	2.9000	6.3000	1.8000	108	.000	.000	1.000	108	.000	.000	1.000
109	6.7000	2.5000	5.8000	1.8000	109	.000	.000	1.000	109	.000	.000	1.000
110	7.2000	3.6000	6.1000	2.5000	110	.000	.000	1.000	110	.000	.000	1.000
111	6.5000	3.2000	5.1000	2.0000	111	.000	.000	1.000	111	.000	.008	.992
112	6.4000	2.7000	5.3000	1.9000	112	.000	.000	1.000	112	.000	.002	.998
113	6.8000	3.0000	5.5000	2.1000	113	.000	.000	1.000	113	.000	.000	1.000
114	5.7000	2.5000	5.0000	2.0000	114	.000	.000	1.000	114	.000	.000	1.000
115	5.8000	2.8000	5.1000	2.4000	115	.000	.000	1.000	115	.000	.000	1.000
116	6.4000	3.2000	5.3000	2.3000	116	.000	.000	1.000	116	.000	.000	1.000
117	6.5000	3.0000	5.5000	1.8000	117	.000	.001	.999	117	.000	.005	.995
118	7.7000	3.8000	6.7000	2.2000	118	.000	.000	1.000	118	.000	.000	1.000
119	7.7000	2.6000	6.9000	2.3000	119	.000	.000	1.000	119	.000	.000	1.000
120	6.0000	2.2000	5.0000	1.5000	120	.000	.000	1.000	120	.000	.302	.698
121	6.9000	3.2000	5.7000	2.3000	121	.000	.000	1.000	121	.000	.000	1.000
122	5.6000	2.8000	4.9000	2.0000	122	.000	.000	1.000	122	.000	.000	1.000
123	7.7000	2.8000	6.7000	2.0000	123	.000	.000	1.000	123	.000	.000	1.000
124	6.3000	2.7000	4.9000	1.8000	124	.000	.000	1.000	124	.000	.094	.906
125	6.7000	3.3000	5.7000	2.1000	125	.000	.000	1.000	125	.000	.000	1.000
126	7.2000	3.2000	6.0000	1.8000	126	.000	.001	.999	126	.000	.003	.997
127	6.2000	2.8000	4.8000	1.8000	127	.000	.000	1.000	127	.000	.162	.838
128	6.1000	3.0000	4.9000	1.8000	128	.000	.006	.994	128	.000	.089	.911
129	6.4000	2.8000	5.6000	2.1000	129	.000	.000	1.000	129	.000	.000	1.000
130	7.2000	3.0000	5.8000	1.6000	130	.000	.005	.995	130	.000	.148	.852
131	7.4000	2.8000	6.1000	1.9000	131	.000	.000	1.000	131	.000	.000	1.000
132	7.9000	3.8000	6.4000	2.0000	132	.000	.006	.994	132	.000	.000	1.000
133	6.4000	2.8000	5.6000	2.2000	133	.000	.000	1.000	133	.000	.000	1.000
134	6.3000	2.8000	5.1000	1.5000	134	.000	.216	.784	134	.000	.746	.254
135	6.1000	2.6000	5.6000	1.4000	135	.000	.000	1.000	135	.000	.073	.927
136	7.7000	3.0000	6.1000	2.3000	136	.000	.000	1.000	136	.000	.000	1.000
137	6.3000	3.4000	5.6000	2.4000	137	.000	.000	1.000	137	.000	.000	1.000
138	6.4000	3.1000	5.5000	1.8000	138	.000	.002	.998	138	.000	.004	.996
139	6.0000	3.0000	4.8000	1.8000	139	.000	.006	.994	139	.000	.123	.877
140	6.9000	3.1000	5.4000	2.1000	140	.000	.000	1.000	140	.000	.001	.999
141	6.7000	3.1000	5.6000	2.4000	141	.000	.000	1.000	141	.000	.000	1.000
142	6.9000	3.1000	5.1000	2.3000	142	.000	.000	1.000	142	.000	.000	1.000
143	5.8000	2.7000	5.1000	1.9000	143	.000	.000	1.000	143	.000	.001	.999
144	6.8000	3.2000	5.9000	2.3000	144	.000	.000	1.000	144	.000	.000	1.000
145	6.7000	3.3000	5.7000	2.5000	145	.000	.000	1.000	145	.000	.000	1.000
146	6.7000	3.0000	5.2000	2.3000	146	.000	.000	1.000	146	.000	.000	1.000
147	6.3000	2.5000	5.0000	1.9000	147	.000	.000	1.000	147	.000	.006	.994
148	6.5000	3.0000	5.2000	2.0000	148	.000	.000	1.000	148	.000	.002	.998
149	6.2000	3.4000	5.4000	2.3000	149	.000	.000	1.000	149	.000	.000	1.000
150	5.9000	3.0000	5.1000	1.8000	150	.000	.002	.998	150	.000	.009	.991

3.1 NORMIX FLOW (MODIFIED) AND OPTIONS

Figure 1, NORMIX flow and options, and Figure 2, NORMIX flow chart, illustrate the general ideas presented above in greater detail. Later, actual data processing of input and the output will provide yet more detail.

In Figure 1, the flow shows in several steps the general course of data information handling. In Figure 2, the flow structure is shown in more detail and the working sections of NORMIX are indicated. The reader is referred to Wolfe's papers. Discussion of these two Figures now follows.

3.2 NORMIX PREPARATION

3.2.1 Input Data Displays

The first box, Input Data Displays, is an option with six suboptions that could be increased at a later date by other users. These options are: 1) chi-square tests, 2) histograms, 3) probability charts, 4) random numbers, 5) scaling, and 6) rotation of wind axes.

3.2.1.1 Chi-square Tests--These ordinary chi-square tests are based on equal probability classes rather than on the usual equal class intervals in terms of the variate values. The null hypothesis is: "the univariate data distribution is not significantly different from a univariate universal normal distribution." A call to an IMSL data package routine is made.

3.2.1.2 Histograms--These bar type charts are in printout form and of the frequencies used in the chi-square tests. They are depicted differently than usual where the bars tend to follow the density curve of the hypothesized model. Instead, if all bar heights are equal and the tops are on a straight horizontal line, the normal model excellently describes the data ensemble for the hypothesized model. A call is made to the IMSL data package.

3.2.1.3 Probability Charts--The probability chart for each variate is yet another look at the potential normality of the variate subset. It is in arithmetic probability form. An upward sloping straight line which would connect the plotted points can be used to examine potential normality of the distribution.

If all of the tests, the chi-square, histograms, and the probability charts, permit the inference that the univariate normal model adequately describe the data sets, the reader must remember that it is the concept of

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
(4-76)

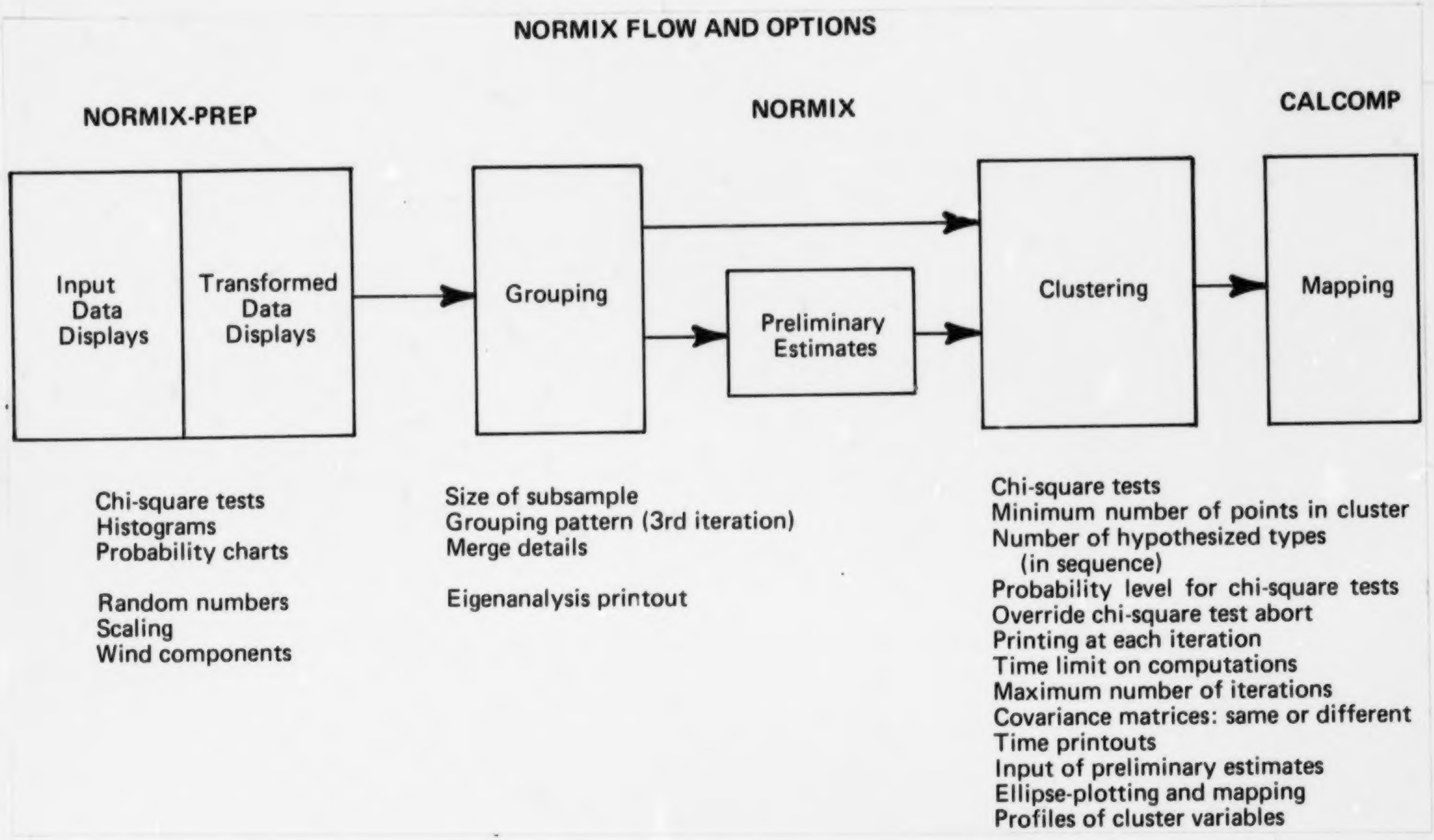

NORMIX FLOW AND OPTIONS
NORMIX-PREP
NORMIX
CALCOMP
Input Data Displays
Transformed Data Displays
Grouping
Preliminary Estimates
Clustering
Mapping
Chi-square tests
Histograms
Probability charts
Random numbers
Scaling
Wind components
Size of subsample
Grouping pattern (3rd iteration)
Merge details
Eigenanalysis printout
Chi-square tests
Minimum number of points in cluster
Number of hypothesized types (in sequence)
Probability level for chi-square tests
Override chi-square test abort
Printing at each iteration
Time limit on computations
Maximum number of iterations
Covariance matrices: same or different
Time printouts
Input of preliminary estimates
Ellipse-plotting and mapping
Profiles of cluster variables

Figure 1. NORMIX flow and options.

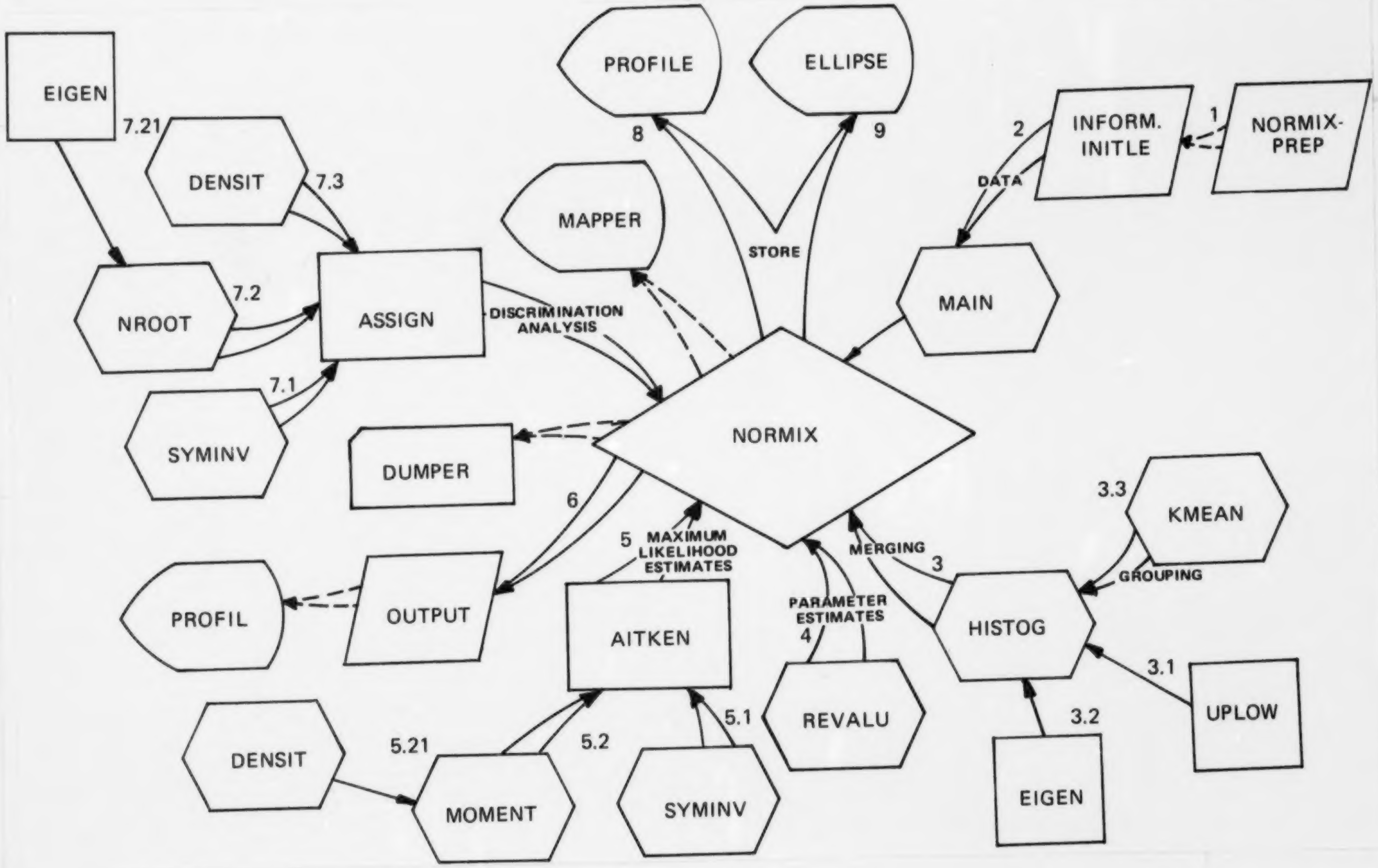

PROFILE
ELLIPSE
8
9
INFORM.
INITLE
1
NORMIX-
PREP
2
DATA
EIGEN
7.21
DENSIT
7.3
MAPPER
STORE
MAIN
NROOT
7.2
ASSIGN
DISCRIMINATION
ANALYSIS
7.1
SYMINV
NORMIX
DUMPER
6
3.3
KMEAN
5
MAXIMUM
LIKELIHOOD
ESTIMATES
MERGING
3
GROUPING
PROFIL
OUTPUT
AITKEN
PARAMETER
ESTIMATES
4
HISTOG
3.1
UPLOW
REVALU
5.1
3.2
DENSIT
5.21
5.2
MOMENT
SYMINV
EIGEN

Figure 2. NORMIX flow chart.

univariate-unimodal normality that is being examined. If the null hypotheses for all the variates concerning unimodal normality are not rejected, the existence of multivariate unimodality is indicated. Therefore, any clustering scheme, including NORMIX, probably will fail to obtain two or more significant clusters.

However, it must be remembered that normality of the single element (variate) marginal data sets does not insure multivariate normality (Crutcher and Falls 1976). That assumption of multivariate normality is made here without testing. Tests are available, including those presented by Crutcher and Falls (op. cit.), but are not included in this program. Such tests could be added later, either as a separate program or as an added option.

If the three options above imply rejection of the null hypotheses that each univariate data set is not significantly different from a normal distribution, then 1) a transformation to approximate normality may be made or 2) the distribution may be multivariate multimodal (hopefully) normal, indicating probable dependable success in clustering. In the case of the multivariate unimodal situation mentioned above, clustering should not be expected beyond the one cluster, the total data set. The decision to transform should be delayed until after the same tests are applied to the output cluster configurations and examined. In this event, a second data run may be made after the transformation. Again, the procedures in some cases are dependent upon the selected transformation. Much care must be used in applying the transformations and in interpreting the results. Some changes in cluster construction and subsequent classification may be noted.

3.2.1.4 Random Numbers--At one time, the early versions of the Wolfe programs included a random number generator option. Somewhere along the line it disappeared. As it appears to be a necessary adjunct, we here provide an option call to the IMSL statistical data package available to the UNIVAC 1110 users at EPA-NCC. It is not an integral part of the preparatory program. At a later date when time is available, a random number generator may be incorporated into the program. Marsaglia, Anathanorayan, and Paul (1972) provide a good generator in the McGill University "Super-Duper" random number generator. The writer has confidence in that particular generator.

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Many sets of data, for one reason or another, may contain a series of constant values for one or more variates. Tests made during the preparation of this program indicate that a constant value for one variate throughout or through one-half of two or more variates will produce singularity problems in the matrix calculations. The existence of such constant values indicates a problem in the validation of the data set for cluster analysis. The investigator must decide whether to reject, modify, or accommodate the data. Accommodation may be effected by use of the random number generator. The option was inserted so that the investigator could modify the constants by adding a small random number of zero mean and variance "V". The program would be able to continue because it would not be halted. Admittedly, this is a crude fix. Knowing that one of the clusters will contain this set of randomized data, the isolation of the set will identify it. Accommodation can also be made in the study of the cluster configuration.

The random number generator can be used also where the singularity problem created by too small a variation might be suspected. Such a case might occur when a sensor or recorder has been dampened.

A multiplier factor of 10^{-5} is in the option, but could be changed easily. However, a change to a larger factor will create more variability, i.e., "V" above will be larger and may change the make-up of the clusters, thereby changing the data classifications. The user need not be too perturbed by these changes, but should study the resultant clusters and classifications within the context of the data and experience. A rerun of the data with a different multiplier factor may be required. A reviewer of this paper does not favor the use of the addition of a random number and recommends the addition of a small multiple of the unit matrix to the within cluster covariance matrix.

3.2.1.5 Scaling--The difference in various variate magnitudes does cause problems. However, they can be met by appropriate scaling. This option permits such scaling.

3.2.1.6 Rotation of wind axes--Generally, wind data are provided in direction in degrees from north and in magnitude in some measure such as kilometers per hour, miles per hour, meters per second, or centimeters per second. Wind

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

speeds follow a central or non-central Rayleighian distribution. Wind direction distributions vary from spike distributions to equal probability distributions that are definitely not normal.

Though wind directions and speeds are not normally distributed, orthogonal components of the wind are normally distributed (Brooks, Durst, and Carruthers 1947, Crutcher 1957). Usually, these concepts are restricted to the troposphere and not to the atmospheric boundary layer. The mixed normal situation is expected to hold within the boundary or mixing layer. However, some investigators are interested in components along specified lines such as highways, coasts, pollution line sources, rivers, etc. An option defaulting to the west-east and south-north permits the user to select any orthogonal set of axes by entering the degrees from north orientation needed. The components will be positive from the south and west and at 180 degrees and 270 degrees clockwise from the option rotation, following meteorological convention.

3.2.2 Transformed Data Displays and Preliminary Estimates

"A priori" information or experience may indicate that the variates individually, in groups, or in totality may not be adequately represented by the normal distribution.

The NORMIX program's maximum likelihood procedures, the convergence problems, and the decision criteria are based on the normality or near normality of the underlying distribution.

There is a serious flaw in many sets of data produced by automatic sensing equipment. Data below minimum detectable limits (BMD) are often recorded. In prior processing, these data may have been deleted or made a constant. These data may also contain inadmissable negative values, although negative values within the noise band are good values in the relative sense. Pollutants, in reality, are always positive, never negative. The recorded negative values pose a problem. The NORMIX program can handle data whether negative or positive, because 1) a logarithm transformation toward normality will be used, and 2) the problem is due to the known existence of BMD data. Again, the NORMIX program would process the data and the results could be used by someone who did know the limitations imposed by the two features, i.e., negativeness and BMD. These particular data can be deleted (rejected), perhaps corrected, modified, accommodated, or set aside for further study. For every negative

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

variable rejected or set aside, an observational vector will be lost. The NORMIX program (as do most programs of this type) requires complete vectors, i.e., when the user specifies in the beginning the number of observations and the number of variates, the computer reads a value for each variable. If a variable or element value is missing, a zero is read and used in the computations. A missing value interpolator could be used. If most of the data are large in value but a good portion of them, say 15 percent, have missing elements, then one cluster will include those which have the missing elements, identified with a zero value with perhaps a variance of zero. The user would wonder why a potential singularity problem may occur.

The procedure used in this report circumvents but does not solve the problem of data with values below a minimum detectable level. Considerable research or just plain simple but continuous redundant monitoring, effective instrumentation calibration, and maintenance is indicated. Here, in this program, an algorithm will determine and add a number to each vector element (variate) to make the least value more than slightly positive. This adjustment is necessary so that the logarithmic transformation models suggested for general use in EPA for pollutants can be applied (Larsen 1971).

Fitz-Simons and Holland (1979) used six models to describe two sets of data. One model is the log normal. The descriptions are compared and suggestions made to better the models. The program they develop is a good and useful tool for the research investigator.

Histograms provided by Fitz-Simons and Holland (1979) indicate the possibility of two subsets or more. These subsets may be real or they may be a sampling feature. Whether they are real or not, does not change the usefulness of the program, which has the acronym MAXFIT. It would be interesting, instructive, and perhaps important to match the data used by Fitz-Simons and Holland (op. cit.) with weather data and/or other pollutant data and to process the matched set through the NORMIX program. The clusters of subsets, if these indicate the same grouping, could then cluster by cluster be subjected to the MAXFIT program processing.

The existence of many BMD data requires the investigator to make some decisions. The data sets often appear to produce a reasonable frequency distribution shape. The instrumental damping or transform may not be known. The circumventing technique may or may not be an adequate transformation. Again, considerable laboratory and instrumental research needs to be done.

The resultant data sets available to the researcher also may be censored or truncated. Some may have all the BMD data made equal to one-half the distance from that level to zero. Some investigators distributed the data equally over the interval. There are arguments pro and con. Creason (1978) has attacked the problem. Mage and Ott (1978) discuss the problems of truncation and censorship. It is recommended that in future work the techniques suggested by Creason (op. cit.) and Mage and Ott (op. cit.) be investigated. At the present, these techniques are limited to one or two variates. In the data sets in this study, there are more than two variates.

The modification of a censored distribution to provide the distribution below or above the censor-points or both is not too difficult in one variate. The problem is discussed in several advanced statistical tests such as Hald (1952), as well as by Creason (op. cit.). However, the treatment of each variate in consonance with the others in a multivariate situation is very difficult. Serial correlations, at the least internal correlations, are involved. Considerable work needs to be done. For example, ozone (O_3) and the oxides of nitrogen (NO_x) amounts are inversely related through the reaction:

$$3NO + O_3 \rightleftarrows 3NO_2.$$

Other pollutants show direct, inverse, or no relationship at all with the above and other pollutants. The position of one pollutant in the BMD region may or may not be related to the position or value of another pollutant or other variate at the same time. The relationship may be with a value above the BMD.

The user is urged to examine carefully the pollutant data distribution prior to transformation, after transformation, and after clustering within each cluster. It is the distributions in the clusters that provide the maximum information on the transformation that perhaps will be needed.

3.2.3 Preliminary Estimates (Hierarchical Grouping)

The NORMIX computer program can handle many variates and a large number

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

of data vectors or observations. At the present time, allowance has been made for 20 variates and 1,000 observations. These numbers can be changed. The main restriction is the cost to reserve space within the computer and the rise in computing time as variates and observations are increased.

It is important to know that preliminary estimates may be inserted into the program for each of the clusters. They may come from previous experience, from clustering output of other programs, or by pure guess. Such preliminary estimates can also be obtained by processing a limited number of observations from the data set through the NORMIX hierarchical clustering and maximum likelihood estimation routines. Wolfe (1967a) suggests that this procedure be done, since close estimates are generally provided, shortening the convergence time required in the maximum likelihood estimation routine of the NORMIX program. This process is called the "learned" or "taught" process, depending upon one's point of view.

If the size of the total sample is small, then all data vectors (observations) may be matched against each other in this initial grouping, leading to estimates for the development of clusters. No difference in grouping will be noted whatever may be the order of data entry. Such is not necessarily the case when the data set is large and initial clustering points are assigned by any method. When the data set is large, the matching or comparing of each pair distance requires an exorbitant expenditure of computer space, time, energy, and money. Therefore, in the NORMIX program, as an option a preliminary number of vector end points up to 150 in number, are identified or denoted as clusters of one observation each, i.e., with 40 vectors or observations, there are 40 initial clusters. These 40 points form the initial points of coalescence for the remaining observational vectors. Thus, final clusters are biased towards the initial points. If the clusters are quite distinct, then the final clustering will be just as distinct no matter what the order of data entry. If the data clusters really overlap, then different cluster estimates may be obtained for the data when they are run in reverse order, random order, or any other order. It is important to remember that the output is only an estimation of group characteristics and that precise duplication of results will only occur with a rerun of the same data in the same order of

9-1/8"
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

entry. It is also important to realize that with other computers with different logic and precision configurations, different arithmetical and even slightly different cluster assignments will be obtained. Local maxima exist in the data configurations. If the initial clustering happens to be in the neighborhood of one of these maxima, then convergence to that local maximum will occur. Just how much difference there is between local and global clustering cannot be determined from a single run on the data. Several runs on the data should be made with different starting data. Random samples can be used to determine likely global maxima. These estimates can then be inserted into the program. An option is provided for this insertion.

In the present modified version of NORMIX, double precision is used throughout. In earlier versions, some parts of the program were in single precision, while other parts were in double precision. This situation created problems in adapting the program to the UNIVAC 1110, although it might not cause problems with other computing systems. Thus, comparison of output among various computer configurations for the same data set will probably differ. Consequently, one program version may not necessarily provide better discrimination and subsequent correct classification. As indicated above, in the maximum likelihood estimates and the convergence process, the order of data entry may permit convergence to be attained at different local maxima rather than the global maximum being sought. Some care is required. If it is known that the clusters may be close together, it is suggested that small blocks of the data be rearranged for entry and two or three runs with different arrangements be made to assure that the local maxima are not creating problems.

3.2.3.1 Initial Clustering Points (Subsample Size)--The size of the subsample is optional. As indicated previously, any number up to 150 can be selected at the present time. If there are more than 150 total data vectors, a default option provides for an initial 150 points assigned as clusters. Forty initial points are suggested, but any number can be chosen. The smaller the subsample, the more chance there is that local maxima problems or different output clusters will be attained.

3.2.3.2 Grouping Patterns (Hierarchical Clustering)--The NORMIX program uses a hierarchical grouping scheme based on the collaboration and work of Ward

EPA-287 (Cin.)
(4-76)

(1963) and Ward, Hall, and Buchhorn (1967). Three iterations of the procedure produce three successive maps, which can be used to form the usual dendritic diagrams. An option calling for only the last of the three maps computed is permitted to save output paper and time.

Wolfe (1967) presents an extensive output of the processing of the Anderson-Fisher (loc. cit.) data to illustrate the procedure. Wolfe (1968) did not present computed cluster ellipses. This present program has been modified to produce secondary data files from which data can be drawn to produce such ellipses. They will be discussed in detail later.

Figures 3a and b, Fisher-Iris data mapping flow, are used to illustrate the third hierarchical mapping of the first 50 data vectors of the Anderson-Fisher (loc. cit.) data; they are the I. setosa data.

Some persistence and patience on the part of the reader is required at this point. Figure 3a illustrates the map as produced by the computer. Please see the schematic on page 3-21 for help.

Fifty original data vectors were available. An initial formation of 40 clusters was called. On the left hand side of the figure are four columns labeled, respectively, RANK, ITEM, KMEANS, and STAGE. The field is forty by fifty, the forty representing the 40 clusters and labeled number of groups which identify the vertical columns. The fifty indicate the set of 50 data vectors, the original order of which is indicated by the ordering numbers, 1-50, in the column RANK. Under the column labeled ITEM, the numbers refer to the original order identification and the negative number indicates that data vector is the first in the formation of any cluster. Thus, ITEM (-1) indicates that the first data vector is the first member of its cluster, namely the first cluster. The fact that it is the first cluster is indicated with a 1 under the K-MEANS column. Note that there are two 1's, indicating that there are two points in the first cluster. The second point is identified by the number 5 under the ITEM column. It is not preceded by a negative sign. Under the STAGE column, the 0 indicates that as yet none of the 40 clusters have merged. The number 41, which is the 40 + 1, indicates an observation of the original 50 that has merged with another. In this case, observation number 5 has merged with observation number 1. This occurrence is further supported by the dot on the first line in column 40 being the point of coalescence for point 5 which is indicated by the x.

```
                                          HIERARCHICAL GROUPING

                                                    NUMBER OF GROUPS
RANK  ITEM  KMEANS  STAGE   1         6         11        16        21        26        31        36
   1    -1       1      D   o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
   2     5       1     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
   3    -8       2     29   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o
   4    28       2     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
   5    40       2     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
   6   -50       3     37   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o
   7   -29       4     39   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o
   8   -49       5     28   X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o
   9    11       5     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  10   -19       6     11   X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  11   -21       7     30   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o
  12    -2       8      8   X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  13    35       8     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  14   -26       9     33   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o
  15   -10      10     23   X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o
  16    13      10     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  17   -12      11     13   X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o
  18    30      11     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  19   -31      12     35   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o
  20   -15      13      4   X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  21   -36      14     14   X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o
  22   -37      15     25   X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o
  23   -42      16      6   X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  24   -25      17      3   X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  25   -45      18     16   X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o
  26   -33      19      5   X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  27   -47      20     26   X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o
  28   -38      21     31   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o
  29   -34      22     18   X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o
  30    -3      23      1   X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  31    -4      24     32   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o
  32    40      24     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  33    -9      25     36   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o
  34    39      25     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  35    -7      26     19   X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o
  36   -43      27     38   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o
  37   -14      28      9   X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  38   -23      29     17   X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o
  39    -6      30      2   X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  40   -20      31     27   X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o
  41   -22      32     24   X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o
  42   -27      33     34   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X o o o o o o
  43   -16      34     12   X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  44   -17      35     20   X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o
  45   -18      36     10   X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  46    41      36     41   X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
  47   -46      37     22   X X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o
  48   -32      38     15   X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o
  49   -44      39      7   X X X X X X X o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o o
  50   -24      40     21   X X X X X X X X X X X X X X X X X X X X X o o o o o o o o o o o o o o o o o o o
```

Figure 3a. Fisher Iris data mapping flow, I. setosa (..., x).

RANK	ITEM	KMEANS	STAGE
1	-1	1	D
2	5	1	41
3	-8	2	29
4	28	2	41
5	40	2	41
6	-50	3	37
7	-29	4	39
8	-49	5	28
9	11	5	41
10	-19	6	11
11	-21	7	30
12	-2	8	8
13	35	8	41
14	-26	9	33
15	-10	10	23
16	13	10	41
17	-12	11	13
18	30	11	41
19	-31	12	35
20	-15	13	4
21	-36	14	14
22	-37	15	25
23	-42	16	6
24	-25	17	3
25	-45	18	16
26	-33	19	5
27	-47	20	26
28	-38	21	31
29	-34	22	18
30	-3	23	1
31	-4	24	32
32	40	24	41
33	-9	25	36
34	39	25	41
35	-7	26	19
36	-43	27	38
37	-14	28	9
38	-23	29	17
39	-6	30	2
40	-20	31	27
41	-22	32	24
42	-27	33	34
43	-16	34	12
44	-17	35	20
45	-18	36	10
46	41	36	41
47	-46	37	22
48	-32	38	15
49	-44	39	7
50	-24	40	21

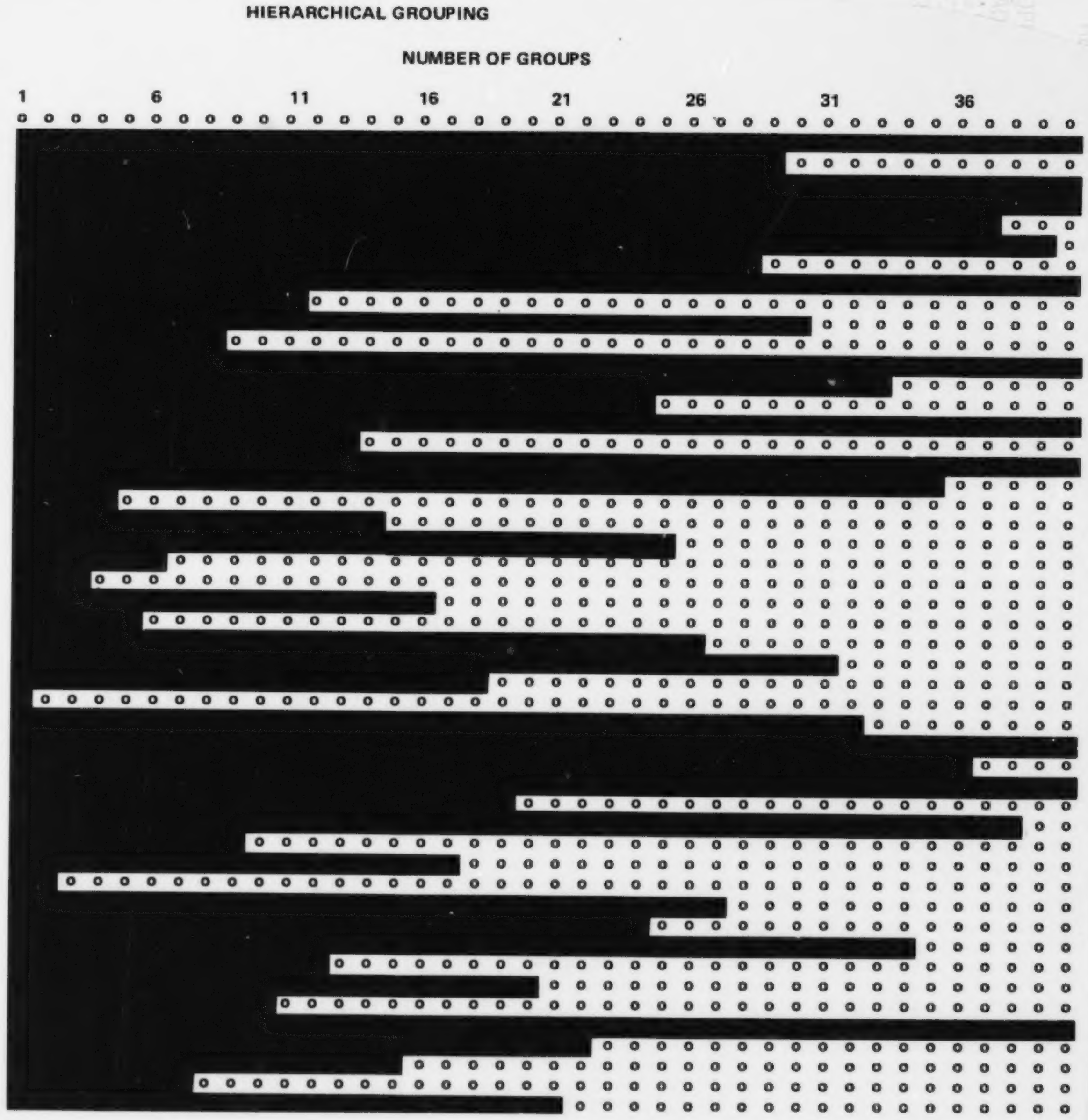


Figure 3b. Fisher Iris data mapping flow, I. setosa (..., ≡).

In Figure 3b, the procedure is shown in more detail. The first 40 clusters are identified in two ways. Under the K-MEANS column, the clusters are identified from 1 to 40 and the make-up of each cluster in terms of the observations is identified under the ITEMS column. In the 40th field column, the 40 cluster structure is available. Each cluster is initiated on the top by a dot. There are 31 singletons, 8 doubletons, and one triplet. The first of these doublets has been identified by the two 1's under the K-MEANS column. The clusters are always identified by the observation having the lower RANK number or order of entry into the computer.

In the 39th column, two singleton clusters of the 40 of the 40th column combine to form a doubleton cluster. They are ITEMS of Rank (-29) and (-50). The bar end in column 39 shows these two points as one cluster. ITEM (-29), the 4th cluster of the forty, has combined with (-50), the 3rd cluster of the forty, under K-MEANS. Under STAGE, the number 39 indicates that there are now 39 clusters. This count checks with the 39th column.

In the 38th column, two more singleton clusters combine, namely clusters 26 and 27, under K-MEANS, and are made up of observations ITEMS (-43) and (-7). Opposite the (-43) under ITEM and under the column STAGE, the number 38 indicates that there are now 38 clusters.

In the 37th column, a triplet and a doublet combine to form a quintuplet. As 39 and 38 cluster formations have already been accounted for, the 37 under the STAGE indicates that there are now 37 clusters. This cluster is composed of vectors 8, 28, 48, 50, and 29.

This procedure is continued. In the 14th column, a triplet is formed by the doubleton formed under the 25th column and the singleton ITEM (-15). In the 7th column, the singleton ITEM (-42) merges with the triplet above just formed. Note that entry has been made into a prior cluster of more than one point, from above as well as from below in the diagram.

Note that the numbers given under STAGE also permit you to better visualize the clustering. For example, the number 1, correlated with the ordered number 30, acting as the third original vector observation entered and as the 23rd cluster (singleton, in this case), is the lead vector of the lower group that joins with the upper group. The lower group is composed of clusters 23 through 40 of the original forty while the upper group is composed of 1

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

through 22 clusters of the first 40 clusters under the K-MEANS column. The makeup of the two clusters may be determined by the identification provided under the ITEMS column.

The following schematic of dots and x's illustrates how to read the joining pattern of singleton vectors and of two clusters:

```
               ..
               ..}  joining of two singletons
               x.
               ..

               ..}  joining of two singletons at bottom
               x.   of mapping table

               ..
               xx}  joining of singleton to cluster above
               x.
               ..

               ..}  joining of singleton to cluster below
               x.
               xx
               xx
               xx

               ..

          xx
          xx
          xx
          x.
          xx     }  joining of two clusters
```

Figure 3b is a modification of Figure 3a, consisting of replacing the x's with solid bars so that the idea of histogram associated with the NORMIX HISTOG may be correlated. The coalescing process can be better seen this way. In all three diagrams, the entry of the last few data into the clusters of more than one point may be seen; just look for two rows of dots that disappear last. For example, the last singleton cluster to combine with another cluster is found at STAGE 6. This cluster is the 42 observation identified in the ITEM column and was the 16th cluster of the original 40. Note that it joins a cluster above in this pictorial diagram. The next to last singleton to enter a cluster is found in the same way. Here, it is found at STAGE 4, identified as cluster 13 of the original 40 and as the 15th original observation of the 50. Note that the last two singletons enter the same cluster, indicating that the singletons and the doubleton which they joined are subjects for examination and discussion as possible outliers to the group remaining of the original 50. However, there is a doubleton which merges with a quadruplet at the fifth STAGE. This cluster is also an outlier candidate in terms of reference of the original 50.

The above discussion indicates that one approach forming data validation after initial screening for impossible data has been made. The NORMIX output also produces corroborative material in a listing of the clustering procedure from which the mapping above is derived. The user may wish to use that instead of the histogram. This output is not shown here. The reader is referred to Wolfe's (op. cit.) papers. The reader has yet to see the construction of cluster ellipses with overprinted classifications which are yet another aid to data validation.

Study of this mapping process from the printout of the clusters, their item contents, and the merging process or from the histogram depiction enables the possible identification of outliers. Thus, this procedure offers good potential in data validation.

Crutcher and Joiner (1977) use these procedures to look at equatorial stratospheric data. Specifically, these are the zonal and meridional components of the wind, the temperatures, and the heights of the pressure level of 30 mbs. Figure 3c taken from that paper and also used by Crutcher schematically shows in dendritic form the coalescence of 74 data points from top to bottom, coalescing from 40 predetermined clusters to one cluster.

For data validation purposes, i.e., situations involved with rogues, outliers, or discordant data, there may be and should be interest in the data vectors, singleton, doubleton, triplets, or even larger group vectors that are the last to join a larger cluster.

3.2.3.3 Merge Details--As another option, the grouping patterns are subsequently also identified by cluster numbers and by the data vectors in each cluster. These merge details are not pictorially represented here. First,

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

the singleton clusters are identified; then the merging process is indicated cluster by cluster. Each new cluster is identified by the cluster having the lower number. Thus, if clusters 17 and 24 are merged, the new resulting cluster is identified as 17. This merging process continues until all data vectors have been gathered into one final cluster, number 1, of course, that is the original data set.

An option permits portions of this merge detail to be printed.

3.2.3.4 Eigenvalue Printout--The eigenvalues and eigenvectors form the basis for matrix orientation and rotation and the determination of Mahalanobis' distances. As an option, they may be printed for examination. In subsequent sections, particularly the quadratic discrimination (classification) and ellipse plotting, other eigenvalue-eigenvector computations are made.

These eigenvectors are the vectors that determine the orientation of the principal axes (variates) with respect to the original axes (variates). Components of the original vectors along these principal axes are not correlated. The eigenvalues are the variances along the principal axes.

In the separation of heterogeneous distributions into more or less homogeneous parts, the multimodality is often clearly shown by the projection of the data vectors onto the main principal axis (Crutcher and Joiner 1977). This projection is not done in the present version of the program output, but could be added as an option by the serious investigator. It should be remembered that the main principal axis has, by design, the largest variance.

3.3 CLUSTERING

Figure 2, NORMIX flow chart, mentioned previously, schematically illustrates the general work flow of the present modified NORMIX program. This flow chart is discussed in much more detail in Volume II of this report series. The flow can be visualized by the numbers arranged in ascending order of operation. Thus "1" is the NORMIX PREP, preparatory section, discussed previously and leads through INFORM-INITLE to a MAIN program leading further to the real central heart of the structure, NORMIX. From this central core, the various operations are controlled until output is obtained and files are prepared for peripheral use, such as plotting of ellipses, profiles, or even other options that may be added later.

EPA-287 (Cin.)
(4-76)

For further details, the user is referred to Wolfe's papers (op. cit.).

Reference is again made to Figure 1. As indicated under the box, "Clustering", many options exist which may be increased or be deleted as time goes on. At the present, these options greatly enhance the program's utility. They are discussed briefly below.

3.3.1 Chi-square Tests

These tests are similar to those of the NORMIX-PREP. As indicated in that discussion, normality, not rejected in the NORMIX-PREP, leads to the assumption of a multivariate-unimodal data set. Clustering beyond the total set will not be expected. Hopefully, if enough data are available in each cluster, the variates within each cluster will be approximately normally distributed. If not, then further breakdown or clustering may be in order.

3.3.2 Histograms

The discussion on the chi-square tests applies here.

3.3.3 Minimum Number of Points in a Cluster

3.3.3.1 Minimum Number of Points--Ordinarily, the minimum number of data points to be clustered in a set is made one more than the number of variates. An option override is provided, permitting data validation or establishment of clusters for data validation for small cluster sizes. Another reason for the override is that many investigators have or can only obtain small samples. The establishment of 2 for a minimum cluster size will then permit the computer to provide some seemingly reasonable clusters.

3.3.3.2 Data Validation--One point in a cluster or an estimated zero point in a cluster creates singularity problems in computation and a cessation of computation for the unequal covariance assumption. Such points may be obvious after detection as key-punch or key-tape errors, malfunctioning equipment, or a variety of error producing sources, and must be rejected or corrected, modified, accommodated or accepted. See the discussion in a prior section. Once these errors have been detected and accounted for, the NORMIX program

BOTTOM OF IMAGE AREA: OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-267 (Cin.)
(4-76)

having done its work in initial screening and data validation can move forward into the field of data or information clustering. Some clusters, proportionately large, may, in themselves, represent data validation and assurance information.

3.3.3.3 Number of Hypothesized Types or Clusters--The program is designed to test (k') vs (k) cluster to determine whether the null hypothesis of no difference is tenable. Generally the difference between (k') and (k) is one. Consecutive tests are made in a series rather than to test all combinations. The number of tests in total would be (k(k-1)/2). Costs to do all of these tests would be excessive.

If a maximum number of clusters, say 10, are to be considered, the entry of numbers 1,2,3,...,10 would initiate testing of 2 vs 1, 3 vs 2, 4 vs 3, ...10 vs 9, unless the procedure is stopped by the non-rejection of a null hypothesis, say 4 vs 3. The user may be interested in the sequence (1, 3, 5, 8, 12). In this case, the testing would procede to examine the null hypothesis for 3 groups versus 1 group, 5 groups versus 3 groups, etc.

3.3.4 Probability Level for Chi-square

The user must choose a decision criterion so that the computer can stop computation or it will continue to test the null hypothesis of (k') vs (k) clusters. Here is an important point: if the chi-square level is exceeded, the requisite cluster computation has already been made. Therefore, this information is also printed. First, it will enable the user to assess the computed probability of rejection. Second, from a practical viewpoint, the user can assess the importance of the cluster information printout. A typical probability level setting is 0.05.

3.3.4.1 Over-ride Chi-square Test Abort--In testing the (k') clusters versus the (k) clusters null hypothesis, non-rejection of the null hypothesis stops the computation. The user may suspect or know that there may be more groups. Thus, though the null hypothesis test of 2 groups versus 1 group may indicate nonrejection, yet further testing of 3 groups versus 2 groups or 1 group may indicate rejection. This rejection happens when three well separated groups have a grand mean equal to or approximately equal to the mean of the central

EPA-287 (Cin.)
(4-76)

group and there is some semblance of symmetry in the cluster configuration. In order to look at all the consecutive tests up to some limit selected by the user, an override option is provided so that the entire sequence of tests wanted will be made. Another course of action is possible. The probability level in (4) above may be changed. The probability level may be set higher, say 0.20 or 0.10, rather than 0.05.

3.3.4.2 Printing at each Iteration--If the user needs for checking purposes an idea as to the iterative process progress, the results of each iteration can be called and printed.

3.3.4.3 Time Limit on Computations--Computation time may be excessive if both the number of variates and data points are large, particularly if the imposed "run time" is surpassed. Provision, therefore, is made to store all data so far processed and the results of such processing up to one minute, or some other arbitrary time in minutes, before the time limit expires for computing. This stored information can be used to again initiate the program at that point and proceed.

3.3.4.4 Maximum Number of Iterations--With singularity problems and on other occasions, excessively slow convergence occurs. At the present time, to keep down excessive computing time, computation ceases at the 100th iteration and the results are printed. This 100th iteration point can be changed without having to recompile the program.

3.3.4.5 Covariance Matrices - Same or Different--The user must make an assumption concerning the covariance matrices. They are the same or different. A call for the first will result in linear discriminant function analysis mapping, classification, and congruent ellipses on each cluster centroid. Assumption of the unequal covariance argument will provide discriminant classification equivalent to quadratic discriminant function analysis classification. Ellipses of different orientations and sizes will be centered on each cluster centroid.

3.3.4.6 Time Print-outs--Time print-outs will be used for each step of the processing so that some experience may be developed for the various parts of the program.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

3.3.4.7 Input of Preliminary Estimates--As discussed briefly before, the computer uses considerable time in the NORMIX-PREP and the grouping and mapping leading to the NORMIX Core. NORMIX can accept other initial estimates. If these estimates are available, time and cost can be reduced. Preliminary runs using only a small portion of the data set will help. As mentioned previously, random or pure guesses may be used.

3.3.4.8 Ellipse Plotting and Mapping--Experience shows that computer output to microfilm, drum plotter or flat-bed plotter can best be attained by immediate output to peripheral storage and later these stored data can be used to produce output for visual interpretation. A CALCOMP routine, modified for this project, allows the plot of up to three probability ellipses for each cluster. For most purposes, it is suggested that no more than one ellipse per cluster be plotted. Choice of the probability levels may be made.

3.3.4.9 Cluster Variate Profiles--Sometimes, for ease of presentation and understanding, it is wise to present information in a form more or less familiar to most people. For example, suppose there are 11 variates: 6 pollutants and 5 meteorological. The data break down into 3 clusters. The means, standard deviations, ranges and/or other information may be plotted on computer peripheral equipment and will be suitable for reproduction. If transformations are used, some care will be required for presentation. This feature will be discussed in some detail later.

3.3.4.10 Time Presentation of Cluster Ellipses--The investigator sometimes may want to pictorially present the clustering in time sequences such as through the day, week, month, or year. Such an option would permit the user to select the number of elliptical presentations which must be adjusted to the same scale. Scaling is influenced by the physical size of the frame allowed by the computer for the physical size of the computer output to microfilm, to the drum plotter, or the flat-bed plotter. No microfilm output is available for this program at the present time.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

4.0 NORMIX CALIBRATION AND OUTPUT

4.1 CALIBRATION

4.1.1 Objective

An objective is to place into the EPA-NCC computer library a good electronic data computer clustering program that can be used for environmental data validation, classification, and/or prediction. The Wolfe (1971b) NORMIX program modified by Crutcher and Joiner (1977) was selected. Further modification has been made during work on this project. Pre-NORMIX data presentation along with the modified NORMIX constitute the program.

4.2 DEMONSTRATION

In order to demonstrate that the program is 1) available, 2) working, and 3) working properly, several sets of data were used:

1. a classical data set composed of the measurements in centimeters made by Anderson (1935) of 150 Iris plants,
2. a set of synthetic data, Wolfe (1971b),
3. a synthetic set of data consisting of three pre-determined three element data sets which could be expanded in size and increasingly separated,
4. a set of maximum-minimum June and July temperatures at the Raleigh-Durham Airport,
5. a set of health related data consisting of trace metals,
6. a set of river discharge data in thousands of cubic feet per second,
7. a set of new filters, 12 impurities, trace metals,
8. a set of new filters, physical and chemical characteristics, and
9. a set of acid precipitation data.

Only the first set will be used here for demonstration. The sets of data 3 through 9 were explored by other investigators either for familiarization with the program or to form part of other research.

4.2.1 Classical Data Set

The classical data set is composed of the measurements made by Anderson (1935) and used by Fisher (1936) in his classic study of the Linear Discriminant Function Analysis (LDF) in "The Use of Multiple Measurements in Taxonomic

Problems". This data set has been used many times by investigators since then, including Wolfe (1967a, 1968). It is useful in the calibration of any clustering program and so is also used here.

The set is comprised of 150 4-variate measurements of Iris plants. There are three subsets of 50 each of I. setosa, I. versicolor, and I. virginica. The four measurements were made on each flower. The measurements were of the sepal length, the sepal width, the petal length, and the petal width.

According to Fisher (op. cit.), the difference between I. setosa and I. versicolor is so great that no appreciable overlapping occurs. There is overlapping of the I. versicolor and I. virginica, perhaps indicating some hybridization and/or polyploidy. Fisher (op. cit.) provides three histograms of the discriminant function analysis output where the regions of intersection can be seen and assessed. Thus, such separation and overlapping features should appear in the output of the now available NORMIX program. For that matter, these same features should be noted in any clustering program. These features will be discussed again. They will be evident in the probabilities associated with each classification and ouput data plots.

Table 2 (page 3-4), in the first four data columns, presents the data gathered by Anderson (1935) and used by Fisher (1936). They were used by Wolfe (1967a, 1968) and are used here for calibration to make sure that the revised program provides essentially the same answer as the results published by Wolfe (1967a, 1968). The first column is the ordering column to which reference may be made in the classification portion. The entries are ordered from 1 to 150. The first 50 represent the I. setosa data, the second 50 represent the I. versicolor data, and the third 50 represent the I. virginica data.

It is stressed once again, that slight differences from the results of different program versions should not bother the investigator too much. Classification should be nearly the same. Differences are expected because of the use of double precision rather than a mixture of single and double precision as in Wolfe (1967a, 1968).

Whenever hierarchial schemes are used to establish initial clusters of which preliminary statistical parameter estimates are made, a change in the order in which data entry is made may produce slightly different clusters. Please refer to pages 3-15 and 3-16. In hierarchial clustering, there is a

EPA-287 (Cin.)
(4-76)

bias towards the first or initial set (subset of data). There is one exception. This exception occurs when each datum of a data set is considered to be a cluster and only distances are used. For example, in the Wolfe NORMIX program, if the hierarchial clustering alqorithm is used to prepare the initial estimates rather than to use the user's initial estimates, different runs on the data, using different data order entry, may produce different preliminary estimates. In the hierarchial clustering algorithm used in the Wolfe NORMIX program ahead of the NORMIX algorithm itself, the user selects an initial subset of the k-clusters of one datum each onto which coalescence of the remaining data is to be made. If there are 40 points (data vectors) in the data set and 40 initial clusters are called then no matter what the order of data entry, the output should not vary from run to run. If there are 150 data points in the data set and 40 initial clusters are called, the order of data entry may be important. The first 40 points are 40 singleton clusters and are points of coalescence. Reversal of data entry order or randomization may produce differing output. Prior ordering of the data in random order is suggested. A limited number of runs suggests that faster and better convergence will be attained by the use of randomly ordered data vectors. Table 2 presents the data in the order published by Fisher (1936) and used by Wolfe (1967a, 1968). The different results will be induced by the different preliminary estimates feeding into the NORMIX algorithm.

Table 2 contains also the output classification produced by the NORMIX for the unequal and the equal covariance assumptions to permit comparison with the Wolfe (op. cit.) output. The order is the same as the input data.

Kendall (1966) applied a pairwise distance clustering technique using the rank order of measurements. He also was able to cluster the entire set of I. setosa with no members of the other two subsets involved. He experienced difficulty in the intersection region of the other two subsets, i.e., with about a dozen of the remaining flowers.

Friedman and Rubin (1967) discuss some of the then available clustering techniques. In considering the relationships of Wilks' Lambda, $T = B + W$, where T, B and W are the "Total", "Between", and "Within" groups sums of squares and cross products, various aspects may be examined with respect to clustering. In using the minimum trace of W criterion, Friedman and Rubin (op. cit.) indicated that I. setosa separated cleanly as one group. A second group was composed mainly of I. versicolor and a third group was composed mainly of I. virginica. The application of the maximization of T / W and trace $W^{-1}B$ criteria recovered three groups except for three plants in both criteria. For the determinant criterion, plants 71 and 84 went to I. virginica and plant 134 went to I. versicolor. For the trace $W^{-1}B$, plants 71, 78, and 84 went to I. virginica. As measured and recorded by Anderson (1935), plants with numbers from 51 to 100 belong to I. versicolor and those with identification 101 to 150 belong to I. virginica.

All of the above discussion and the following discussion assume, of course, that the measurements made by Anderson (op. cit.) were without error and that his grouping is without error. The inference is made also that the data published by Anderson (op. cit.) and used by Fisher (op. cit.) were the same as Anderson (op. cit.) obtained. Anderson's (1935) grouping may have been in error because of the existence of real rogues created by seed dispersal. Presumably, no one will ever know the real answers, but the assumption must be made that no errors exist although clustering techniques may seem to isolate and classify certain plants, repeatedly, differently than the original classification. Whether this situation is so is not germane to the testing of a revised version of the NORMIX program against the original program. Nevertheless, it is an interesting and provocative feature and will be mentioned again.

Wolfe (1967a, 1968) uses two assumptions: equal covariance matrices and unequal covariance matrices. Under the first assumption of equal covariance matrices, plants 71 and 84 are classified as I. virginica with probabilities of 0.867 and 0.873, respectively, rather than as I. versicolor, while plant 134 assumed to be I. virginica is classified as I. versicolor with a probability of 0.746. The output of the present program provides the same results. These results are the same as those cited by Friedman and Rubin (1967) for the determinant criterion.

Under the assumption of unequal covariance matrices, Wolfe (1967a) indicates considerable departure in classification, though again, plant 71 was classified as I. virginica rather than I. versicolor. Plants from the I. virginica classification of Anderson (1935) were classified as I. versicolor, namely, plants 104, 106, 108, 109, 117, 118, 120, 123, 126, 130, 131, 132, 135, and 138. There is a difference between the Wolfe (1967a) program and the output of the revised program now available under the assumption of unequal covariance matrices. However, it is noted that about the same number of plants gave Wolfe (1967a) and Kendall (1966) difficulty in the intersection region.

The program now being calibrated classifies plants 69, 71, 73, 78, and 84 as I. virginica rather than the assumed I. versicolor group. Plants 71 and 84 are changed from I. versicolor type to I. virginica type with probabilities of 0.947 and 0.993, respectively, which are greater, and therefore, may be better than the 0.867 and 0.873 previously indicated. Plant 134 retains its original classification but the probabilities are 0.874 versus 0.216, so there is still some indication of indecision though the ratio is near 4 to 1.

The above comparisons and calibration are interesting. The assumption of equal covariance matrices provides fairly good agreement. The assumption of unequal matrices reveals differences in the intersection region. The agreements are sufficiently close to permit the assumption that the revised version now available for the NORMIX program is somewhat better than the 1967 version. The differences are not so great as to imply that calibration is unacceptable. As only 5 plants were reclassified by the now available NORMIX and two of these were reclassified by all procedures discussed, the applicability of the present NORMIX routine is judged to be better. Double precision, as used in the revised version, may permit the changes in classification. Further work or investigation of the reclassification of plants 71 and 84 may reveal the reasons for the shift. That research is beyond the scope of the present work.

4.2.2 Synthetic Data Set

Wolfe (1965) constructed a set of data with predetermined characteristics. These synthetic data were used to calibrate his NORMIX program. User comments, including those made during the preparation of the paper by Crutcher and Joiner (1977) were used to revise Wolfe's 1967a, 1968, and 1971b program. Since then, users have requested and received the modified version. That version of the NORMIX program has been placed on the IBM 360-165 computer at TUCC. A run made on that computer using the Wolfe (1971) synthetic data set show differences in computed classification probabilities, but only one different classification was created by the 1974 revision. The output of the TUCC computer and the 1110 EPA-NCC computer after processing the same set of data shows complete agreement to three decimals in all data checked on a random sample check.

4.2.3 Other Data Sets

The remaining data sets 3 through 9, mentioned previously, were processed on the EPA-NCC computer. The results seem to be acceptable but can only be shown to be so by the individual investigators. Clustering was accomplished in most of the sets.

4.3 RESULTS OF CALIBRATION

Calibration is considered to be complete and acceptable. The Wolfe (1967a, 1968) program, now modified, is available and working at the EPA-NCC complex.

4.3.1 Output

Table 2, previously discussed, provides the classical set of Iris data and the resultant classifications obtained through the use of the two assumptions, equal covariance matrices and unequal covariance matrices.

4.3.1.1 Intermediate Output--The intermediate cutput is illustrated now in a very abbreviated form.

Normality Checks, Univariate-Unimodal--Table 3 provides 1) the chi-square, 2) histogram, and 3) arithmetic probability computer output for the first variable, sepal length. In the sense of a unimodal univariate normal model the chi-square test and a visual scan of the histogram indicates multimodality of variate 1. The arithmetic probability chart does not appear to deviate much from the expected upward sloping straight line of a unimodal univariate normal distribution. There are two possible decisions. If the data are univariate normal the data are either unimodal or multimodal. Any indication that the single mode may not be a tenable assumption leaves only the assumption of multimodality. The inference is made that the distribution is sufficiently multi-modal to permit clustering. The results of the clustering overall, however, do not allow this particular variate to separate by itself, but it undoubtedly influences the separation when all four variables are used. The probability of obtaining a Chi-square value larger than our calculated value is only 0.0000742, assuming the null hypothesis is true.

4.3.1.2 Core Output--Table 4 provides the cluster estimates derived from processing the initial estimates provided by the hierarchical clustering scheme. They are the maximum likelihood estimates. The input data are 150, the number of variates is four, and the number of cluster types being tested against previous computations (not shown) for 2 cluster types are 3. Twenty-two iterations were required in the solution of the maximum likelihood equations. The inference is made that 3 groups are significantly different than the previous 2 groups. The characteristics of the total sample and each of the approximated cluster types are then presented. The estimates are of the proportions, the means, the standard deviations, and the correlations. They are used in the classification procedures which produced the second part of Table 2. These estimates also go into the secondary data files from which they may be extracted to prepare plots of ellipse and profiles.

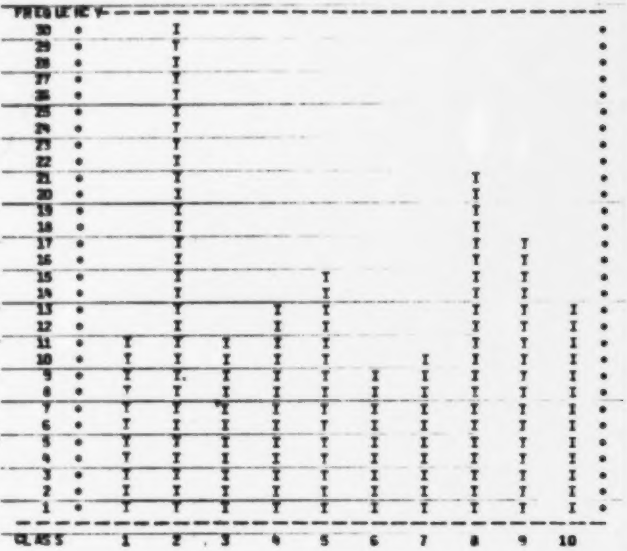
4.3.1.3 Post NORMIX--Figure 1 shows several options for output data. Two sets are selected for presentation here.

Normality Checks, Univariate-Unimodal--The normality checks 1) chi-square, 2) histogram, and 3) arithmetic probability chart presentation of calculations of sepal length were presented. The normal distribution univariate and unimodal was rejected, opening the possibility of clustering as several modes may be inferred from a visual scan of the histogram.

For Type 1, I. setosa, Table 5 provides the chi-square test for univariate unimodality and a histogram for visual inspection. The chi-square test does not indicate rejection of the null hypothesis. However, from inspection of the histogram chart, one can infer that two groups do exist in these data, which leads to inference that in subsequent finer clustering the I. setosa would break into two groups. This process does not happen though in the first five clusters unless logarithmic transformation of the sepal lengths is made. However, it is not the purpose here to make a detailed study of the Iris data except as it applies to the calibration of the program's Wolfe's NORMIX core

Table 3. Fisher Iris data; 150 sepal lengths distributed in equal probability classes with chi-square frequency table and test, histogram, and arithmetic probability chart

```
                                                      ANALYSIS OF VARIABLE  1
MEAN VALUE                                         =   5.8433
STANDARD DEVIATION                                 =    .8281

CELL NUMBER     FREQUENCY
     1              11.
     2              30.
     3              11.
     4              13.
     5              15.
     6               9.
     7              10.
     8              21.
     9              17.
    10              13.
CHI-SQUARE                                         =  24.4000
PROB. OF EXCEEDING CHI-SQUARE IF NULL HYP. IS TRUE =    .0010
DEGREES OF FREEDOM                                 =   7
```



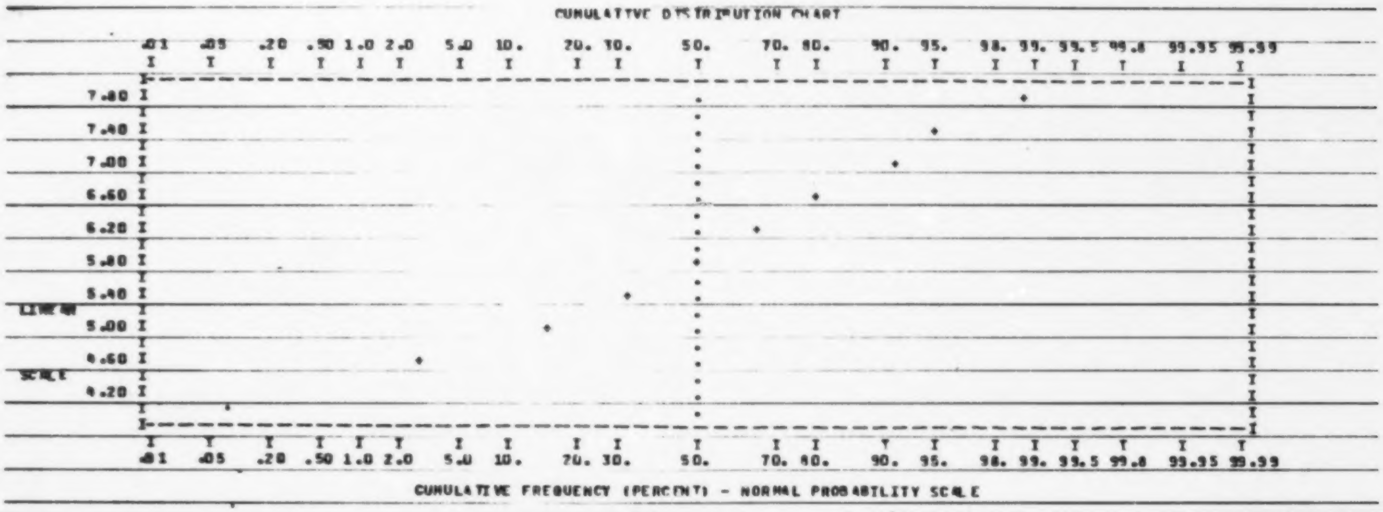


Table 4. Three cluster output of Fisher Iris data; 150 input data, proportions, means, standard deviations, correlation matrices, and null hypothesis test

```
                                  PROGRAM NORMIX
                    WOLFE NORMAL MIXTURF ANALYSIS PROCFDURF(1979 RFVISION)

NORMIX PROGRAM WITH IRIS DATA

      SAMPLE SIZE =          150
      NUMBER OF VARIABLFS =    4
      NUMBER OF TYPFS =        3

                     ITERATION NUMBFR  22

   LOG LIKELIHOOD OF  3 TYPES IN THIS SAMPLE =      .37117763+003

        CHARACTERISTICS OF THE WHOLE SAMPLE

 SEP LNG      SEP WTH      PET LNG      PET WTH
     CM           CM           CM           CM
                                                             MEANS
    1            2            3            4
 5.8433       3.0573       3.7580       1.1993
                                                     STANDARD DEVIATIONS
    1            2            3            4
  .8253        .4344       1.7594        .7597
                                                        CORRELATIONS
    1            2            3            4
 1.0000       -.1176        .8718        .8179
 -.1176       1.0000       -.4284       -.3661
  .8718       -.4284       1.0000        .9629
  .8179       -.3661        .9629       1.0000

                 CHARACTERISTICS OF TYPE   1

          THE PROPORTION OF THE POPULATION FROM THIS TYPE=  .333

 SEP LNG      SEP WTH      PET LNG      PET WTH
     CM           CM           CM           CM
                                                             MEANS
    1            2            3            4
 5.0060       3.4280       1.4620        .2460
                                                     STANDARD DEVIATIONS
    1            2            3            4
  .3489        .3753        .1719        .1043
                                                        CORRELATIONS
    1            2            3            4
 1.0000        .7425        .2672        .2781
  .7425       1.0000        .1777        .2328
  .2672        .1777       1.0000        .3316
  .2781        .2328        .3316       1.0000

                     CHARACTERISTICS OF TYPE   2

  THE PROPORTION OF THE POPULATION FROM THTS TYPE=  .299

   SEP LNG      SEP WTH      PET LNG      PET WTH
       CM           CM           CM           CM

      1            2            3            4
   5.9144       2.7778       4.2016       1.2970

      1            2            3            4
    .5247        .3044        .4479        .1789

      1            2            3            4
   1.0000        .6069        .7857        .5795
    .6069       1.0000        .6685        .7897
    .7857        .6685       1.0000        .7611
    .5795        .7897        .7611       1.0000
                        CHARACTERISTICS OF TYPE   3

  THE PROPORTION OF THE POPULATION FROM THTS TYPF=  .367

   SEP LNG      SEP WTH      PET LNG      PET WIH
       CM           CM           CM           CM

      1            2            3            4
   6.5446       2.9487       5.4796       1.9847

      1            2            3            4
    .6221        .3322        .5725        .2929

      1            2            3            4
   1.0000        .4462        .0502        .3383
    .4462       1.0000        .4432        .5757
    .0502        .4432       1.0000        .4443
    .3383        .5757        .4443       1.0000
```

Table 5. Fisher Iris data; I. setosa, 50 sepal lengths in equal probability class intervals, chi-square frequency table and test, histogram

```
 THE FOLLOWING CHI SQUARE TEST AND EQUAL PROBABILITY  HISTOGRAM ARE FOR TYPE NUMBER  1 AND VARIABLE NUMBER  1
CELL NUMBER     1    2    3    4    5    6    7    8    9   10
FREQUENCY       5.   6.   5.   4.   8.   0.   8.   4.   5.   5.

CHI-SQUARE =    9.2000    PROB. OF EXCEEDING CHI-SQUARE IF NULL HYPOTHESIS IS TRUE =    .2386    DF =    7
STANDARD DEVIATION OF THIS VARIABLE =     .3525

 FREQUENCY-------------------------------------------
     8  *                   I       I               *
     7  *                   I       I               *
     6  *       I           I       I               *
     5  *   I   I   I       I       I       I   I   *
     4  *   I   I   I   I   I       I   I   I   I   *
     3  *   I   I   I   I   I       I   I   I   I   *
     2  *   I   I   I   I   I       I   I   I   I   *
     1  *   I   I   I   I   I       I   I   I   I   *
 ----------------------------------------------------
  CLASS     1   2   3   4   5   6   7   8   9  10
```

program. Further study is reserved for those who want to do so. Clustering at times is sensitive to data transformations. Transformations must be used with care.

Ellipse Output--An objective was to produce visual output to enable research workers to visualize the clustering effect. The representations chosen here are two dimensional ellipses. In the future, three-dimensional ellipsoids will be possible after programs are completed and added as an option.

The variate pair correlations and the standard deviations are used here to compute the eigenvector-eigenvalue combination required for the construction of the ellipses. Crutcher (1957) describes the technique. Figures 4, 5, and 6 provide 0.30 probability ellipses in the 2-space of petal length versus petal width of the Anderson (op. cit.) and Fisher (op. cit.) Iris data. Figures 4, 5, and 6, respectively, present 3, 4, and 5 clusters.

Again, it is pointed out that the number of paired variates permitting plotting is equal to (n(n-1)/2) where n is the number of variates. Thus, for each cluster, in the case of 21 variates, there will be 210 possible bivariate presentations. If there are 1, 2, and 3 cluster configurations, then there are 1260 possible bivariate pairs. The user is urged to make a visual scan of the computer tabular output before choosing the pairs for plotting. Principally for that reason, the secondary data files were established.

Examination of the chosen ellipses of the bivariate pairs, petal length, and petal width, show that throughout five clusters, I. setosa remains one group. I. versicolor and I. virginica break into two, then three, and then four clusters. The viewer can also assess the potential for further clustering.

The cluster means of the two variates are the centroids of the respective ellipses. The eigenvector and eigenvalues for each variate pair are the principal axes and the variances along the respective axes. The major principal axis is paired with the largest eigenvalue or variance. Crutcher (1957) among others shows how to compute these values and how to obtain the lengths of axes for selected probability levels. Within the present program, three probability ellipses may be computed for each cluster. A default option will produce only the 0.50 probability ellipse.

IRIS DATA N=150

VARIABLE	CLUSTER 1	CLUSTER 2	CLUSTER 3
	P= .333	P= .299	P= .367
SEPAL LENGTH(CM)			
MEAN	5.01	5.92	6.54
STANDARD DEVIATION	0.35	0.52	0.62
SEPAL WIDTH(CM)			
MEAN	3.43	2.78	2.95
STANDARD DEVIATION	0.38	0.30	0.33
PETAL LENGTH(CM)			
MEAN	1.46	4.20	5.48
STANDARD DEVIATION	0.17	0.45	0.57
PETAL WIDTH(CM)			
MEAN	0.25	1.30	1.98
STANDARD DEVIATION	0.10	0.18	0.29

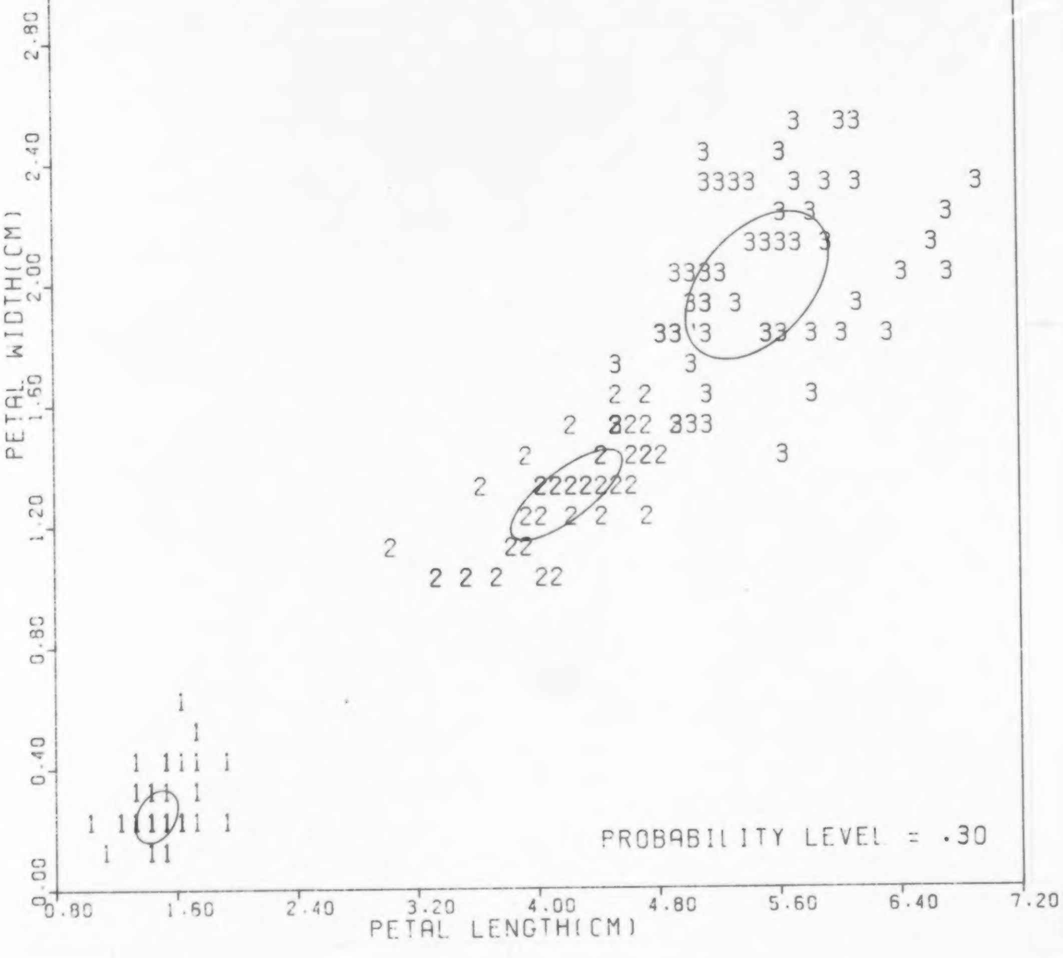


Figure 4. Ellipses, 0.30 probability, 3 clusters, petal length versus petal width, Fisher Iris data.

IRIS DATA

VARIABLE	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
SEPAL LENGTH (CM)	P= .333	P= .298	P= .309	P= .059
MEAN	5.01	5.91	6.36	7.52
STANDARD DEVIATION	0.35	0.52	0.48	0.24
SEPAL WIDTH (CM)				
MEAN	3.43	2.78	2.91	3.10
STANDARD DEVIATION	0.38	0.30	0.31	0.41
PETAL LENGTH (CM)				
MEAN	1.46	4.20	5.30	6.39
STANDARD DEVIATION	0.17	0.45	0.42	0.35
PETAL WIDTH (CM)				
MEAN	0.25	1.30	1.99	1.97
STANDARD DEVIATION	0.10	0.18	0.31	0.21

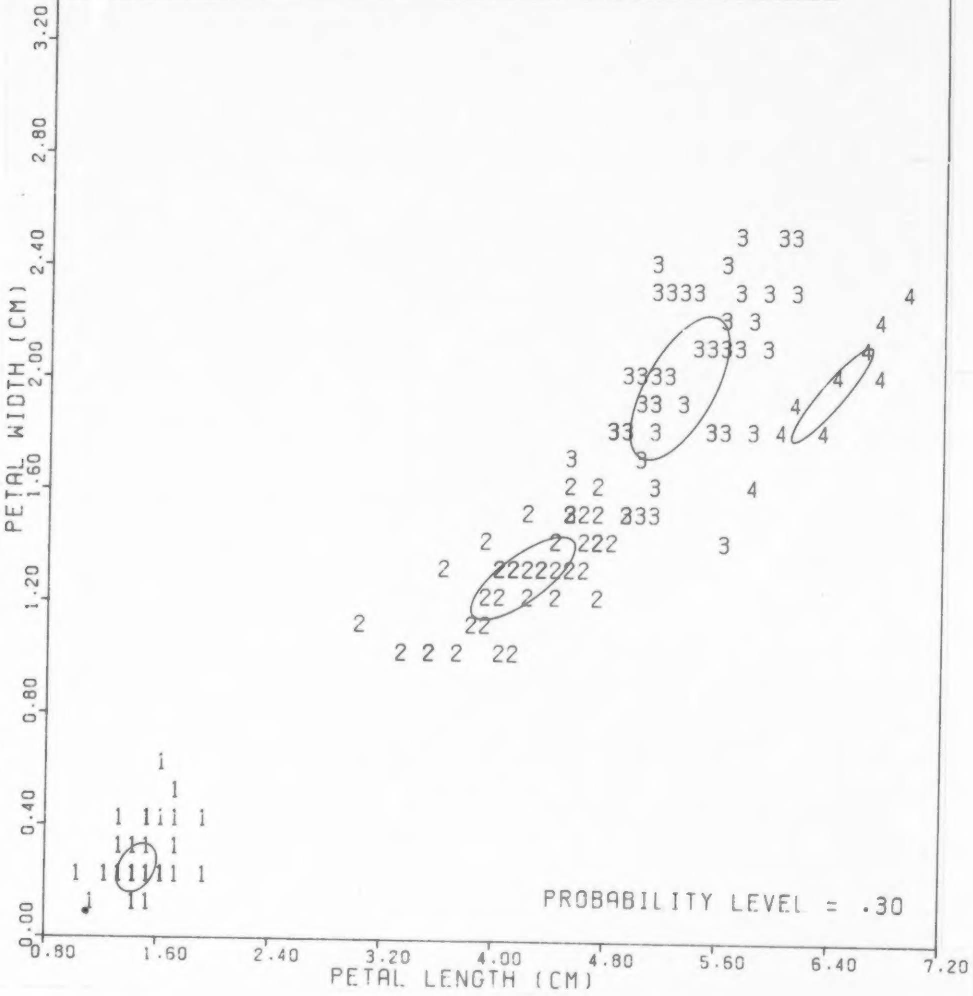


Figure 5. Ellipses, 0.30 probability, 4 clusters, petal length versus petal width, Fisher Iris data.

IRIS DATA N=150

VARIABLE	CLUSTER 1	2	3	4	5
	P= .333	P= .230	P= .083	P= .294	P= .059
SEPAL LENGTH(CM)					
MEAN	5.01	5.69	6.57	6.37	7.53
STANDARD DEVIATION	0.35	0.37	0.27	0.48	0.24
SEPAL WIDTH(CM)					
MEAN	3.43	2.75	2.81	2.93	3.10
STANDARD DEVIATION	0.38	0.30	0.34	0.29	0.41
PETAL LENGTH(CM)					
MEAN	1.46	4.09	4.56	5.34	6.39
STANDARD DEVIATION	0.17	0.44	0.25	0.40	0.35
PETAL WIDTH(CM)					
MEAN	0.25	1.28	1.40	2.01	1.97
STANDARD DEVIATION	0.10	0.19	0.15	0.30	0.21

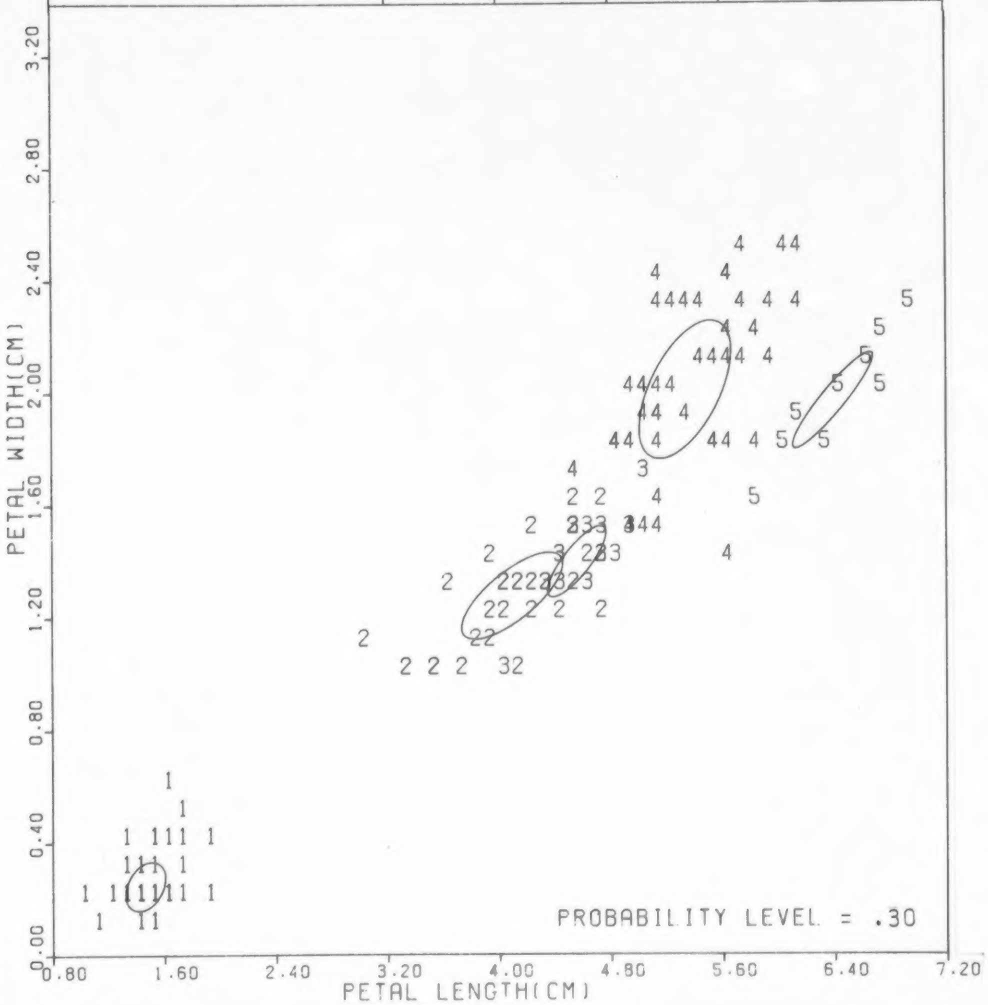


Figure 6. Ellipses, 0.30 probability, 5 clusters, petal length versus petal width, Fisher Iris data.

Any pair of variates may be used for the two dimensional field. Plotted on this field are data points located by the two variates and identified by the cluster type to which the multivariate point has been assigned. Thus, the assignment is multivariate but the plot is bivariate, amounting in essence to a type of projection onto the bivariate plane. A different assignment might be made if only the two variates were used for clustering and discrimination. The ellipses are then drawn over the plotted points in the bivariate plane.

If the assumption of equal covariance matrices has been made, all ellipses (of the same probability level) will have the same size and orientation centered at the various cluster centroids. If the assumption of unequal covariance matrices has been made, the ellipses may have different sizes and orientation in these presentations. If a point in a cluster is duplicated, the replication of the point may not be seen. If the same value is attained by two different cluster types, the overprinting of the designators may perhaps be seen.

The output has been designed so that the proportions, means, and standard deviations are shown for each plot. The addition of the covariance or correlation matrices would require too much space. Therefore, the user should later insert the information wanted within the frame of the plotted output figures.

Scaling--Scaling is perhaps one of the most difficult features involved in the production of plotter output. The physical dimensions of the medium chosen as well as the variates' units must be considered. In the construction of a set of ellipses for several clusters within the same frame, the range both vertically and horizontally must be computed and coordinated with the ranges of the actual data if data plots or scattergrams are to be made. Then these groups must be scaled to the computer and media. The media are cathode ray output to microfilm, drum, flat-bed, or other types of plotters. Additional complexities are introduced when several frames of ellipses are to be presented in a vertical, horizontal, or matrix form. The scaling must be consistent, frame to frame, so that meaningful visual comparisons may be made.

Profiles--Data output results are most easily compared when there is standardization. Again, the scaling problem arises and so does the back transformation if transformation has been used. In terms of quality assurance and control for data validation, the usual configuration of the mean plus and minus three sigma limits are used for each variable. This configuration is used for this profile model. The variates can be presented in the same order as in data entry, or the order of the data entry can be opted to force a rising curve for the largest cluster.

The grand mean for each variable is located on the central mean line. In terms of the physical constraints, the positions of the means are located on the central horizontal grand mean line. If transformations to normality have been used, then it is the grand mean of the transformed values that is used for each variate. If the logarithmic transformation has been used, then this value is the geometric mean $\bar{x}_g$ that is plotted on the central line. Aitchison and Brown (1976) provide a good discussion on the lognormal distribution, as does Larsen (1971). If the reciprocals are used, the mean would be the harmonic mean. It has not been used here in this report. However, if the user employs this transformation, then the mean could be identified as $\bar{x}_h$. The program does provide a power transformation option. The user will have to provide a suitable symbol for this such as $\bar{x}_p$ where p is the power ranging from minus infinity to plus infinity, the infinity being excluded. Moore (1957) discusses such power transformations.

Options are provided to compute selected percentiles of the transformed data and then to backtransform to the scale of the original variable units. An option permits the user to obtain a tabular output. All or some of this information may be plotted in terms of the transformed data, yet can be identified in terms of the original units. In the case of the logarithmic transformation, the mean of the logarithms plotted on the central grand mean line will be identified with a value equivalent to the median of the original values. In similar fashion, the means of the cluster variables will be scaled to the distribution of the total data set but will be plotted and identified with a value equivalent to the median of the cluster variable. The plus two and minus two sigma points will be determined in terms of the transformed variable, plotted in terms of the scale of the total data set.

Where no transformation has been used, all information plotted is in terms of the original variable. These procedures are used so that people other than the immediate user will not be confounded by transformation problems. It is stressed here that though there will be symmetry in the plotted points using the transformed values, there will be no symmetry in the values printed in the table corresponding to plotted points.

As noted in the previous discussion, several sets of data were processed to assess the working potential of the NORMIX program. One of these sets involved the physical and chemical characteristics of new, unused filters. These characteristics or variates were (1) retention, (2) tensile strength, (3) thickness, (4) flow, (5) integrity, (6) and pH. The identifiers (1) through (6) are carried through in Figures 7 and 8. Neither the original data nor NORMIX computations are shown here. Only the profile presentations for the two and then the three cluster configurations are shown, respectively, in Figures 7 and 8. One group on both figures is identified as the group containing the proportion 0.318. The group containing the remaining proportion 0.682 in Figure 7 breaks down into two clusters containing (by coincidence) 0.318 and 0.364. The information given in Figures 7 and 8 allows the means and standard deviations to be extracted for each variate in each cluster. The correlations are not available. No further identification is made as to shipment number, manufacturer, or analysis techniques, etc., at this point as the entire exercise was to examine the efficacy of the NORMIX program to cluster data. Quite evident in Figure 8 are the differences exhibited by the first group 0.318, which remains unchanged from Figure 7 and the group, 0.364. The relationship among the variates are inverse except for the integrity, which remains essentially unchanged. The reader can make other comparisons.

Here, are two helpful hints. The first is that the proportions given in the clustering output are estimates and counts of classified data may not agree with these estimated proportions. This should be no cause for alarm.

BOTTOM OF IMAGE AREA OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT HERE
3/8"
PAGE NUMBER

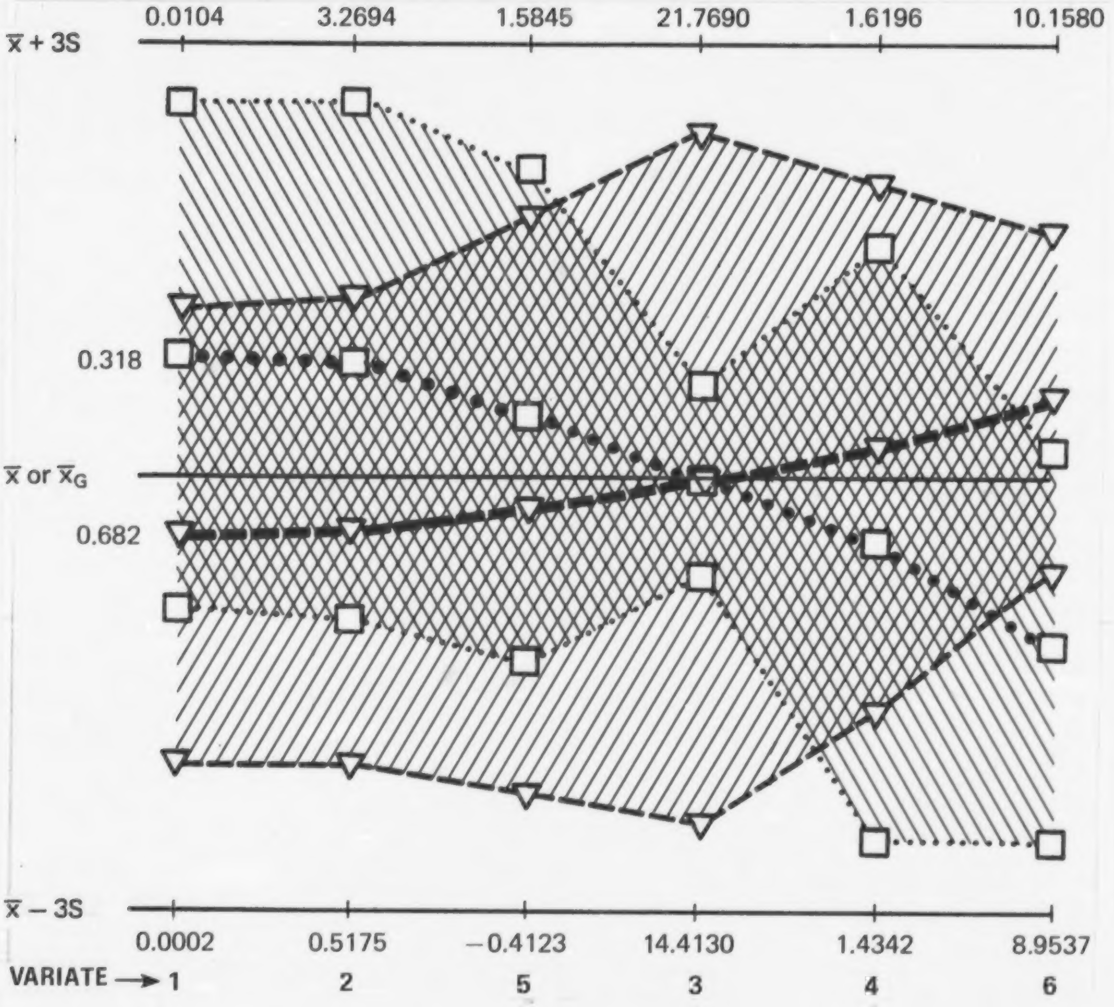


Figure 7. Profile plots for the total, and two cluster data for new filters, physical and chemical characteristics: (1) penetration (%), (2) tensile strength (kg), (3) thickness (mils), (4) flow (cmm), and (5) integrity (mg), and (6) pH.

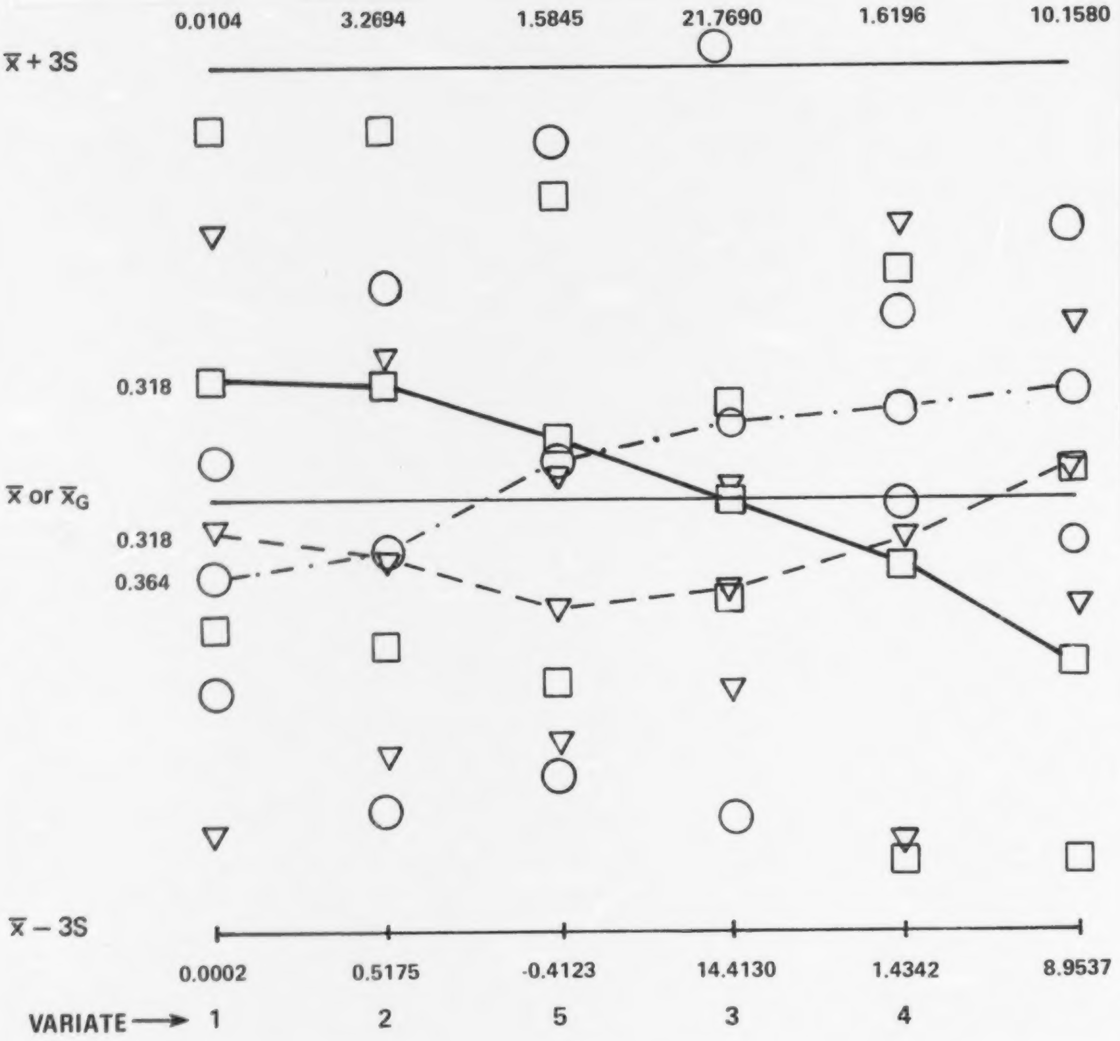


Figure 8. Profile plots for the total, and three cluster data for new filters, physical and chemical characteristics: (1) penetration (%), (2) tensile strength (kg), (3) thickness (mils), (4) flow (cmm), and (5) integrity (mg), and (6) pH.

The second is more serious. In the Post-NORMIX options, these work well when the distances between modes are equal to or greater than six times the standard deviations. For Cluster 1 of the Iris data there is no problem. However, application of the post option analysis to clusters 2 and 3 will be difficult. There are problems of censoring, truncation and folding involved which are beyond the scope of this present project. For the present, the user is advised to be careful when using the post option analyses in their current form.

5.0 PROCESSING OF SELECTED CHAMP DATA

CHAMP is the acronym for Community Health Air Monitoring Program. Three locations were selected with the idea that one is known to be relatively "clean," one "dirty," and one "very dirty." It is stressed here that these terms are relative insofar as the atmospheric pollutants are concerned, the point being that differences in pollutant levels can be expected. The selected locations were at 1) Angwin, California, 2) Magna, Utah, and 3) Loma Linda, California. Location descriptions were kindly furnished by Mr. José Suñe whose (1978) paper may be examined with respect to the Utah area.

5.1 DATA PROBLEM EXAMPLES

Figure 9 illustrates the hourly averages of various atmospheric pollutants and wind information at Loma Linda, CA. This result is version 6.15 of VALDAT for Julian Day 103 1977. Note that some data are missing while other data are negative. The wind speed, the vector wind mean, and the vector wind direction are the meteorological variables available in this format. Some of the contaminants (variates) are the oxides of nitrogen (NO_x); (calculated) nitrogen oxide (NO); nitrogen dioxide (NO_2); sample NO (SNO); ozone (O_3); sulfur dioxide (SO_2); total hydrocarbons (THC); and nonmethane hydrocarbons (NMHC). Temperature, dew points, winds, and atmospheric pressure were obtained separately, if available.

The NORMIX program will accept and use negative values. However, measurements of pollutants should always be positive, never negative. However, they may be essentially zero. In actual measurement, negative values occur within the noise band. Where this situation occurs and the negative variables are to be used with a logarithmic transformation to approximate normality, they must first be made positive.

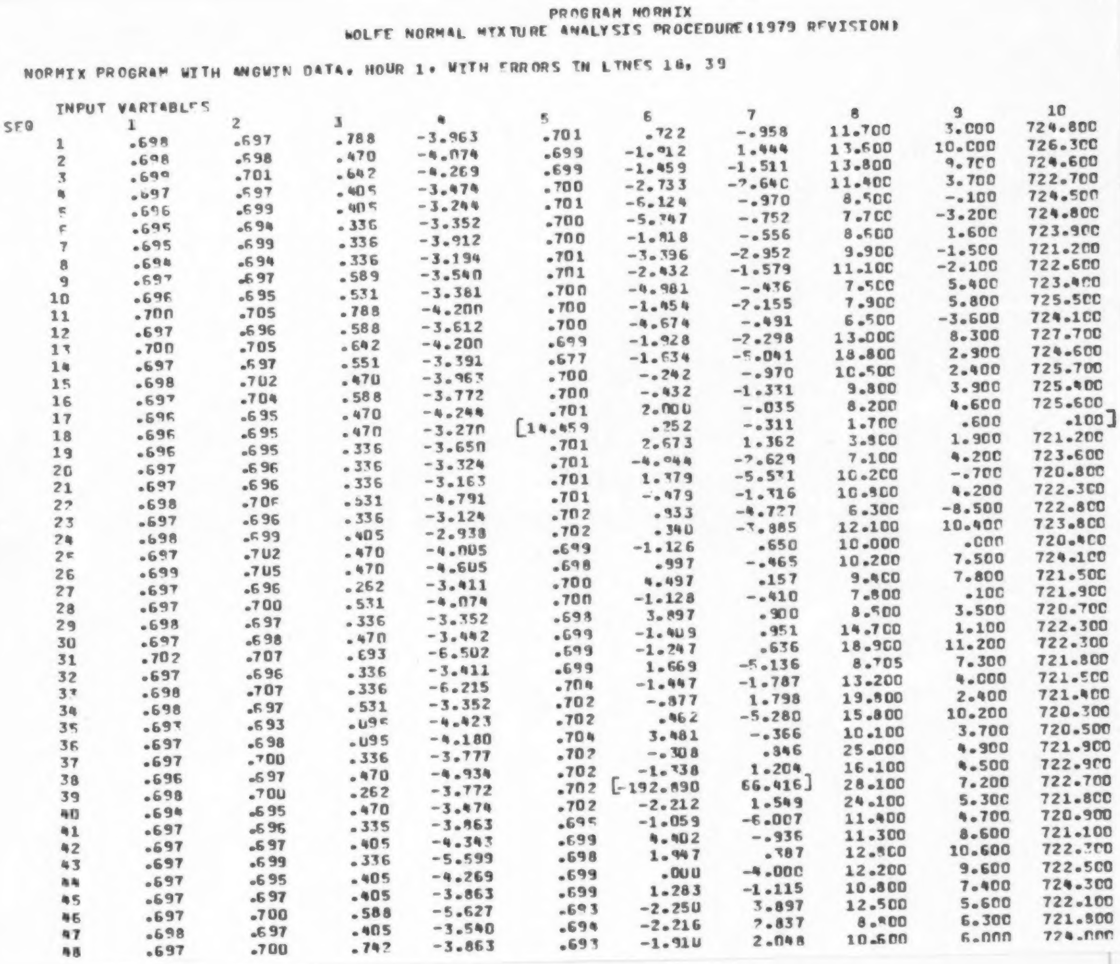
Figure 10 illustrates the occurrence of data which at first scan appear to be "rogues." These data were prepared manually by key-punching from Angwin, CA, hour 1, data records. The data in lines 18 and 39 are so outstanding that a check revealed that these were key-punch errors and were corrected.

TYPING GUIDE SHEET
CENTER
BEGIN FIRST LINE OF TEXT HERE
DROPPED HEAD. BEGIN SECTIONS HERE
TOP OF IMAGE AREA
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"

```
          CHAMP DATA,VALIDATED ON 14-MAY-77 WITH VERSION 6.15 OF VALDAT

     VALIDATION REPORT FOR STATION - 0823  FOR DAY - 103 - 1977

              HOURLY AVERAGES

TIME      NOX        NO        NO2        SNO        O3         SO2        CH4       THC       NMHC       WS        VWM       VWD
 0: 0   0.1105 12 0.0295 12 0.0811 12 0.0831 12-0.0040 12 0.0043 12   3.05 12   4.54 12   1.88I11   1.7 12    0.2 12    0.2 12
 1: 0   0.1111 12 0.0351 12 0.0760 12 0.0775 12-0.0040 12 0.0038 12   3.00 12   3.74 12   0.87I11   2.1 12    0.9 12    3.3 12
 2: 0   0.1100 12 0.0354 12 0.0746 12 0.0765 12-0.0040 12 0.0033 12   2.66 12   2.86 12   0.47I10   2.0 12    0.8 12  174.6 12
 3: 0   0.1225 12 0.0558 12 0.0666 12 0.0676 12-0.0040 12 0.0033 12   3.03 12   4.20 12   0.73I 9   2.1 12    1.1 12  330.5 12
 4: 0   0.1147 12 0.0522 12 0.0625 12 0.0639 12-0.0040 12 0.0032 12   3.56 12   4.27 12   0.99I10   2.4 12    1.0 12  295.0 12
 5: 0   0.1102 12 0.0463 12 0.0639 12 0.0651 12-0.0037 12 0.0033 12   3.44 12   4.43 12   1.36I10   2.4 12    0.9 12  235.1 12
 6: 0   0.1698 12 0.1059 12 0.0639 12 0.0653 12-0.0027 12 0.0047 12   2.57 12   3.21 12   0.59 12   1.7 12    0.8 12   47.4 12
 7: 0   0.0997M11 0.0374M11 0.0623M11 0.0632M11 0.0082M11 0.0052M11   2.14M11   2.40M11   0.21M11   3.4M11    2.3M11   83.4M11
 8: 0   0.1044M 8 0.0307M 8 0.0737M 8 0.0750M 8 0.0208M 8 0.0057M 8   2.11M 8   2.38M 8   0.23M 8   4.3M 8    4.0M 8   77.2M 8
 9: 0   0.0791M 9 0.0091M 9 0.0700M 9 0.0713M 9 0.0539M 9 0.0066M 9   2.10B 7   2.38B 7   0.24B 7   4.8M 9    3.7M 9  119.5M 9
10: 0   0.0586 12 0.0019 12 0.0568 12 0.0582 12 0.0987 12 0.0072M11   2.59M 4   2.20M 4   0.24B 3   6.5 12    4.3 12  351.2 12
11: 0   0.0384 12 0.0001 12 0.0383 12 0.0395 12 0.0926 12        M 0   2.01 12   2.27 12   0.23 12   7.3 12    2.2 12  284.2 12
12: 0   0.0437M 9 0.0009M 9 0.0428M 9 0.0438M 9 0.0984M 9 0.0060B 7   1.89M 8   2.00M 8   0.08M 8  11.1M 9    7.4M 9  321.8M 9
13: 0   0.0350M11-0.0028M11 0.0378M11 0.0386M11 0.1196M11 0.0262 12   1.77 12   1.91 12   0.12 12  16.1 12   14.4 12  293.8 12
14: 0   0.0264B 1 0.0033B 1 0.0231B 1 0.0236B 1 0.1041B 8 0.0066 12   1.67 12   1.75 12   0.06 12  12.2 12    9.9 12  295.7 12
15: 0   0.0224M11 0.0038M11 0.0187M11 0.0188M11 0.0995B 4 0.0060M11   1.70M11   1.78M11   0.06M11  13.3M11    9.0M11  301.4M11
16: 0   0.0246M 9 0.0044M 9 0.0202M 9 0.0206M 9 0.0984M 4 0.0050M 9   1.62M 9   1.69M 9   0.05M 9  13.8M 9   12.0M 9  274.9M 9
17: 0   0.0252M 8 0.0038M 8 0.0215M 8 0.0219M 8 0.0795M 8 0.0048M 8   1.66M 8   1.73M 8   0.08B 7   5.5M 8    3.9M 8  278.6M 8
18: 0   0.0183M11 0.0010M11 0.0173M11 0.0176M11 0.0718M11 0.0045M11   1.57M11   1.60M11   0.01M11   6.9M11    5.2M11  254.0M11
19: 0   0.0153M11 0.0006M11 0.0147M11 0.0149M11 0.0623M11 0.0051M11   1.81M11   1.88M11   0.05M11   5.3M11    3.5M11  306.2M11
20: 0   0.0112 12 0.0007 12 0.0105 12 0.0105 12 0.0620 12 0.0051 12   1.81 12   1.87 12   0.06I11   7.0 12    5.0 12  284.2 12
21: 0   0.0102 12 0.0004 12 0.0099 12 0.0099 12 0.0630 12 0.0051 12   1.75 12   1.67 12   0.02I 8   3.4 12    3.1 12   23.1 12
22: 0   0.0115 12 0.0004 12 0.0111 12 0.0111 12 0.0617 12 0.0051 12   1.62 12   1.65 12   0.01 12   4.0 12    3.1 12   45.1 12
23: 0   0.0096 12 0.0009 12 0.0087 12 0.0088 12 0.0647 12 0.0049 12   1.76 12   1.84 12   0.05 12   4.0 12    3.4 12   27.7 12
```

Figure 9. CHAMP data example for Loma Linda, California, for day 103, 1977, with some indicated bad data, Station Identifier 0823.

TYPING GUIDE SHEET
CENTER
BEGIN FIRST LINE OF TEXT HERE
DROPPED HEAD; BEGIN SECTIONS HERE
TOP OF IMAGE AREA
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER



```
                                 PROGRAM NORMIX
                 WOLFE NORMAL MIXTURE ANALYSIS PROCEDURE(1979 REVISION)

NORMIX PROGRAM WITH ANGWIN DATA, HOUR 1, WITH ERRORS IN LINES 18, 39

     INPUT VARIABLES
SEQ     1       2       3        4        5         6        7        8        9       10
  1   .698    .697    .788   -3.963     .701      .722    -.958   11.700    3.000  724.800
  2   .698    .698    .470   -4.074     .699    -1.912    1.444   13.600   10.000  726.300
  3   .699    .701    .642   -4.269     .699    -1.459   -1.511   13.800    9.700  724.600
  4   .697    .697    .405   -3.474     .700    -2.733   -2.640   11.400    3.700  722.700
  5   .696    .699    .405   -3.244     .701    -6.124    -.970    8.500    -.100  724.500
  6   .695    .694    .336   -3.352     .700    -5.747    -.752    7.700   -3.200  724.800
  7   .695    .699    .336   -3.912     .700    -1.818    -.556    8.600    1.600  723.900
  8   .694    .694    .336   -3.194     .701    -3.396   -2.952    9.900   -1.500  721.200
  9   .697    .697    .589   -3.540     .701    -2.432   -1.579   11.100   -2.100  722.600
 10   .696    .695    .531   -3.381     .700    -4.981    -.436    7.500    5.400  723.400
 11   .700    .705    .788   -4.200     .700    -1.454   -2.155    7.900    5.800  725.500
 12   .697    .696    .588   -3.612     .700    -4.674    -.491    6.500   -3.600  724.100
 13   .700    .705    .642   -4.200     .699    -1.928   -2.298   13.000    8.300  727.700
 14   .697    .697    .551   -3.391     .677    -1.634   -5.041   18.800    2.900  724.600
 15   .698    .702    .470   -3.963     .700     -.242    -.970   10.500    2.400  725.700
 16   .697    .704    .588   -3.772     .700     -.432   -1.331    9.800    3.900  725.400
 17   .696    .695    .470   -4.244     .701     2.000    -.035    8.200    4.600  725.600
 18   .696    .695    .470   -3.270  [14.459      .252    -.311    1.700     .600     .100]
 19   .696    .695    .336   -3.650     .701     2.673    1.362    3.900    1.900  721.200
 20   .697    .696    .336   -3.324     .701    -4.944   -2.629    7.100    4.200  723.600
 21   .697    .696    .336   -3.163     .701     1.379   -5.571   10.200    -.700  720.800
 22   .698    .706    .531   -4.791     .701     -.479   -1.316   10.900    4.200  722.300
 23   .697    .696    .336   -3.124     .702      .933   -4.727    6.300   -8.500  722.800
 24   .698    .699    .405   -2.938     .702      .340   -3.885   12.100   10.400  723.800
 25   .697    .702    .470   -4.005     .699    -1.126     .650   10.000     .000  720.400
 26   .699    .705    .470   -4.605     .698      .997    -.465   10.200    7.500  724.100
 27   .697    .696    .262   -3.411     .700     4.497     .157    9.400    7.800  721.500
 28   .697    .700    .531   -4.074     .700    -1.128    -.410    7.800     .100  721.900
 29   .698    .697    .336   -3.352     .698     3.897     .900    8.500    3.500  720.700
 30   .697    .698    .470   -3.442     .699    -1.409     .951   14.700    1.100  722.300
 31   .702    .707    .693   -6.502     .699    -1.247     .636   18.900   11.200  722.300
 32   .697    .696    .336   -3.411     .699     1.669   -5.136    8.705    7.300  721.800
 33   .698    .707    .336   -6.215     .704    -1.447   -1.787   13.200    4.000  721.500
 34   .698    .697    .531   -3.352     .702     -.877    1.798   19.800    2.400  721.400
 35   .693    .693    .095   -4.423     .702      .462   -5.280   15.800   10.200  720.300
 36   .697    .698    .095   -4.180     .704     3.481    -.366   10.100    3.700  720.500
 37   .697    .700    .336   -3.777     .702     -.308     .346   25.000    4.900  721.900
 38   .696    .697    .470   -4.934     .702    -1.338    1.204   16.100    4.500  722.900
 39   .698    .700    .262   -3.772     .702  [-192.890  66.416]  28.100    7.200  722.700
 40   .694    .695    .470   -3.474     .702    -2.212    1.549   24.100    5.300  721.800
 41   .697    .696    .335   -3.863     .695    -1.059   -6.007   11.400    4.700  720.900
 42   .697    .697    .405   -4.343     .699     4.402    -.936   11.300    8.600  721.100
 43   .697    .699    .336   -5.599     .698     1.947     .387   12.900   10.600  722.300
 44   .697    .695    .405   -4.269     .699      .000   -4.000   12.200    9.600  722.500
 45   .697    .697    .405   -3.863     .699     1.283   -1.115   10.800    7.400  724.300
 46   .697    .700    .588   -5.627     .693    -2.250    3.897   12.500    5.600  722.100
 47   .698    .697    .405   -3.540     .694    -2.216    2.837    8.400    6.300  721.800
 48   .697    .700    .742   -3.863     .693    -1.910    2.048   10.600    6.000  724.000
```

Figure 10. CHAMP data at Angwin, California, hour 1, showing key punch errors in lines 18 and 39, Station Identifier 0812.

EPA-287 (Cin.)
(4-76)

If the above "rogues" were to be in the data entered into the NORMIX programs, and if clusters of less than 11, i.e., the number of variables plus one were allowed, the singularities induced would isolate these "rogues" and stop further computations.

Another problem that may exist in some data sets is the occurrence of constant values inserted by some preprocessing algorithm for BMD values. These values may create collinearity problems in the NORMIX processing. In this case, the option for the random number generator may have to be used. Later modifications to the program through the use of additional algorithms may resolve the problem.

Still another problem is created by missing variables, i.e., missing observational vector elements. The computer will enter a zero and continue to process the data, constituting an error that may cause singularity problems in computation.

A precheck serves a useful purpose to eliminate the above potential troublesome data problems. Of course, the NORMIX computer program will detect most of these, but it remains a misuse of the program.

5.2 STATION LOCATION DESCRIPTIONS

5.2.1 Angwin, California

The Angwin station number in the CHAMP network is 0812. The station is located on the campus of the Pacific Union College at Angwin, CA. The latitude and longitude are 38 degrees, 34 minutes, and 09 seconds north latitude, and 122 degrees, 26 minutes, and 34 seconds west longitude. Please see the attached map, Figure 11.

More specifically the station is located on a site between the center and the western edge of the campus on a parking lot. The parking lot is restricted to day-time employees of the University's computer center.

This parking lot was chosen because it was found to be the only area on the campus that offered the least amount of air disturbance due to traffic movement. Most vehicles park on the lot area between 7:30 and 8:00 a.m. and remain parked until 4:30 to 5:00 p.m. The campus is situated on a hillside and the only flat areas for locating a CHAMP station were found on roadways and parking lots.

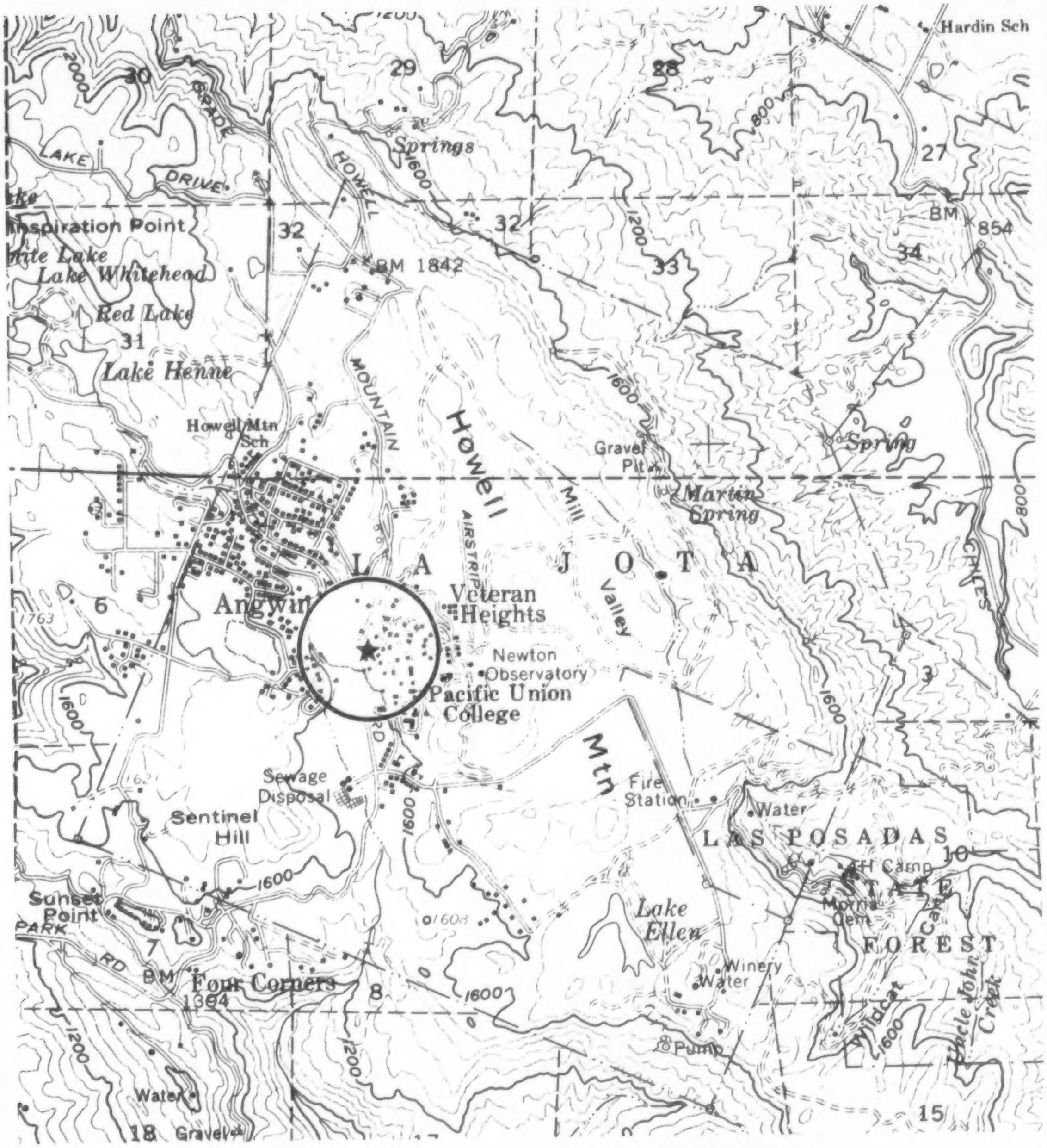

Hardin Sch
30
29
28
27
2000
GRADE
LAKE
DRIVE
HOWELL
Springs
1600
800
Inspiration Point
Lake Whitehead
Red Lake
31
Lake Henne
32
BM 1842
32
33
34
BM
854
1200
Howell Mtn Sch
MOUNTAIN
Howell
Mill
Valley
AIRSTRIP
LA JOTA
Gravel Pit
Spring
Martin Spring
Angwin
6
1763
Veteran Heights
Newton Observatory
Pacific Union College
Mtn
Sewage Disposal
Sentinel Hill
Fire Station
Water
LAS POSADAS
STATE
FOREST
4H Camp
10
Morris Cem
3
Sunset Point
PARK
RD
7
BM
1394
Four Corners
8
1608
Lake Ellen
Winery
Water
Pump
Wildcat Can
Uncle John Creek
15
Water
18
Gravel

Figure 11. Map showing the location of the Angwin, California, CHAMP Station 0812.

The site is flanked on three sides by buildings as follows: Approximately 12 meters to the east and north, there are two two-story buildings with a clear space between buildings approximately 10 meters long. About 25 meters to the west there is a four-story building housing the computer center. The entrance to the parking lot is located 20 meters to the south. This site is almost unobstructed except for low trees and shrubbery. The buildings are far enough away to allow the air-intake a full 45 degree clearance. The air intake probe is approximately 3.5 meters above ground. The altitude above mean sea level is 553 meters.

5.2.2 Magna, Utah

The Magna CHAMP site is located at Magna, Utah. Its number in the CHAMP network is 0541. The site is located at 8560 W 29355 South Street. The latitude and longitude are, respectively, approximately 40 degrees, 42 minutes, and 20 seconds north latitude, and 112 degrees, 05 minutes, and 40 seconds west longitude. Please see the attached map, Figure 12 for more detail.

The station is located along the south edge of the local high school track field. The high school building is located approximately 100 meters to the northwest. There is another school yard immediately to the south where the Magna Junior High school is located, approximately 150 meters from the station. The nearest residence is located approximately 30 meters directly east. To the north there is an open view of the copper smelter (Kenecott), located about 6 kilometers away.

5.2.3 Loma Linda, California

The CHAMP station 0823 at Loma Linda, California, is located on the campus of the Loma Linda University, La Sierra campus at Riverside, California. The latitude and longitude, respectively, are approximately 33 degrees, 54 minutes, and 35 seconds north latitude, and 117 degrees, 30 minutes, and 00 seconds west longitude. Please see the attached map, Figure 13.

The station is located near the approximate center of the campus where most student activities take place. The area is a clear grassy area close to a pedestrian walk and close to a special parking lot restricted to day parking. The nearest building is approximately 30 meters away (College Hall), a

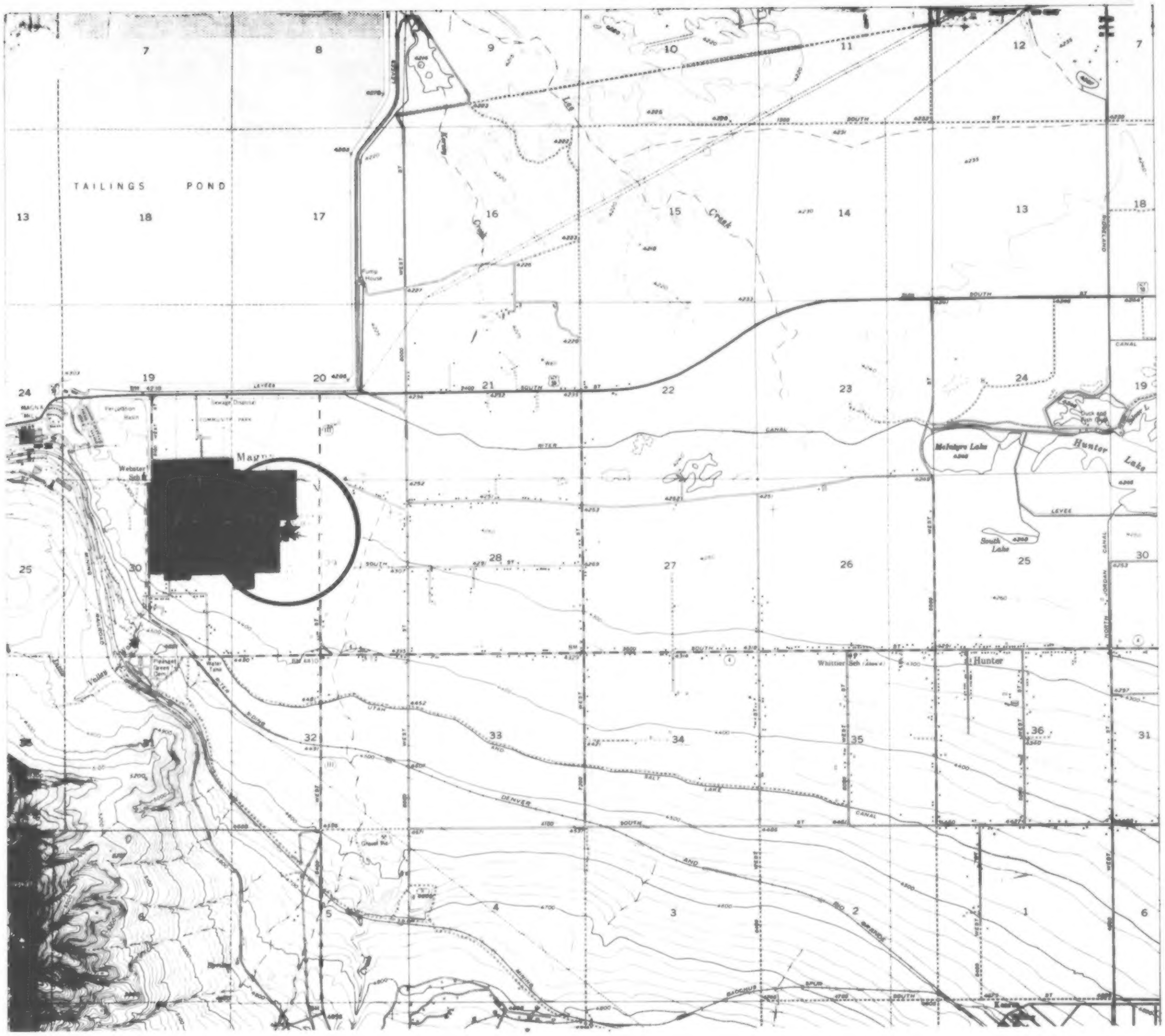

TAILINGS POND
Magna
Webster Sch
Hunter
Melodpro Lake
South Lake
Hunter Lake
Whittier Sch
CANAL
LEVEE
DENVER AND SALT LAKE
SOUTH
WEST
ST
7
8
9
10
11
12
13
18
17
16
15
14
24
19
20
21
22
23
25
30
28
27
26
32
33
34
35
36
31
5
4
3
2
1
6

Figure 12. Map showing the location of the Magna, Utah, CHAMP Station 0541.

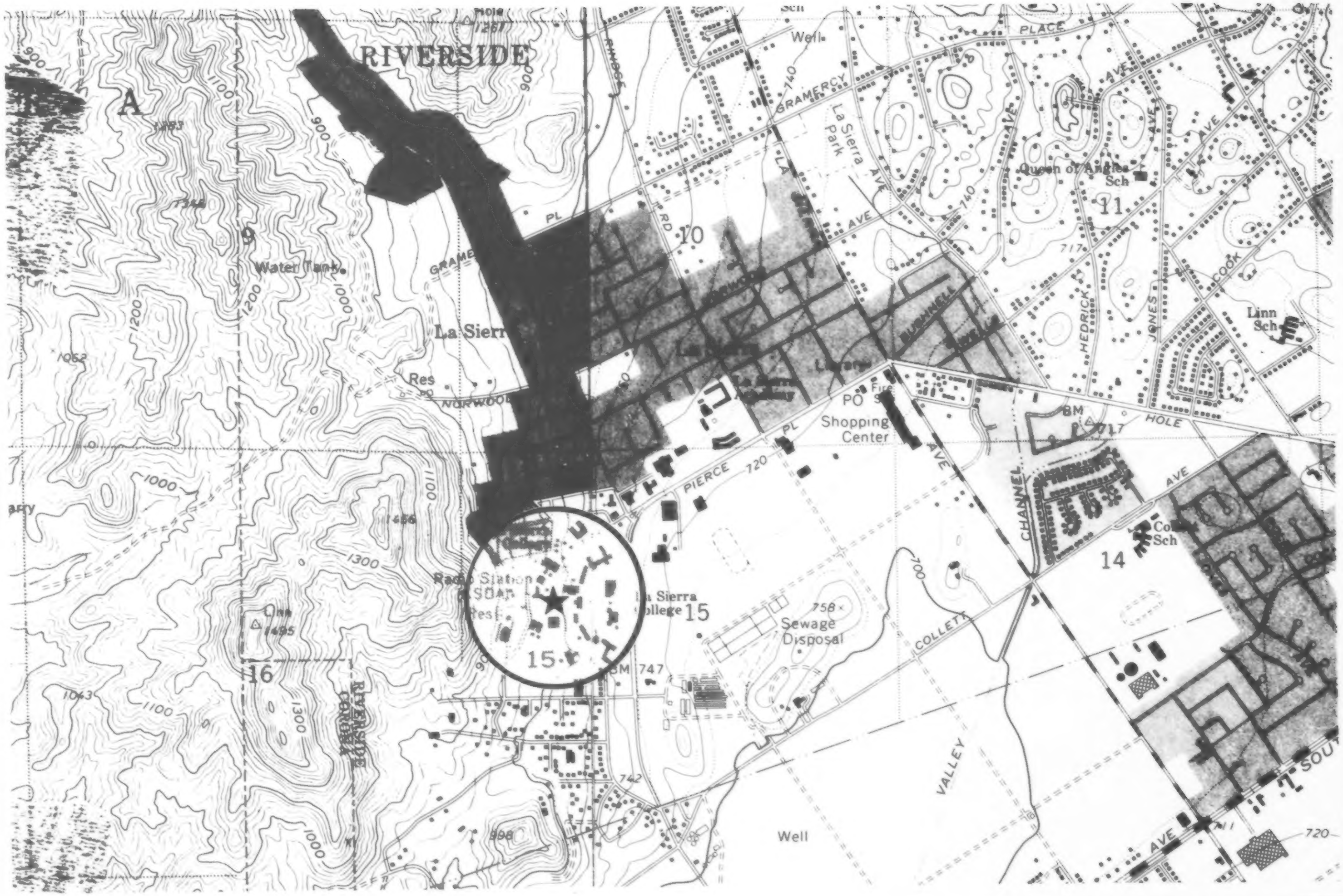

RIVERSIDE
A
9
10
11
14
15
16
Water Tank
La Sierra
Res
NORWOOD
GRAMERCY
La Sierra Park
Queen of Angels Sch
Linn Sch
HEDRICK
JONES
COOK
HOLE
BM 717
PO
Shopping Center
PIERCE
PL
AVE
CHANNEL
Co Sch
Radio Station
La Sierra College
BM 747
Sewage Disposal
COLLETT
VALLEY
Well
RIVERSIDE CORONA
1300
1100
1000
1200
900
700
720
740
758
742
1495
1456
1043
1062
998

Figure 13. Map showing the location of the Loma Linda, California, CHAMP Station 0823.

one-story gymnasium-auditorium. The air intake to the sampling equipment at this station is 3.5 meters from the ground. The altitude of the site is 256 meters above mean sea level.

5.3 SELECTION OF DATA FOR PROCESSING

The total problem is extensive. To determine whether there would be clustering of atmospheric pollutant data and the concomitant meteorological data, the decision was made to selectively sample over the hours of the days, the days of the weeks, the weeks of the years, the weeks of the seasons, two seasons in the year, (summer and winter), and to combine the days' data for an annual summary.

It was realized that the very action of sampling through these periods might induce grouping or clustering. It was also realized that with larger sets of data, better clustering is probable. Many sets and subsets of data were processed and examined. Availability of data in some areas did control the number of variates examined, the hours of the day, and other aspects of the sampling.

Table 6 is a summary of significant clusters obtained by NORMIX processing of selected CHAMP data at Angwin, California, Magna, Utah, and Loma Linda, California. The significance probability level to test the null hypothesis that k'-clusters are not different from k-clusters is 0.05. With computed probabilities less than 0.05, the null hypothesis is rejected.

As this study is on clustering, it is instructive that the clusters be obtained within the selected hours, days of the week, the weeks, and the seasons. The hours are in local standard time with 01 being 1:00 a.m. The days of the week are 1,2,3, and 4, respectively, Sunday, Tuesday, Thursday and Saturday. The weeks of the year are approximately mid-quarter weeks starting with the first week of the calendar year. These are weeks 7, 20, 33, and 46. The seasons were summer and winter in the usual sense, combining the data from the appropriate weeks. The number of variates varied but are not identified at the moment. The number of observations in each case are given in Table 6. The numbers are not large. The period of record at the CHAMP stations is not long: namely Angwin, California (9/75-11/76), Magna, Utah (8/74-9/76), and Loma Linda, California (9/75-11/76). Angwin, California, the cleanest station, produces by hour of the day and by week the fewest clusters while it

Table 6. Summary of Significant Clusters obtained by the use of the NORMIX Computing Program for selected periods of CHAMP Data at Angwin, California, Magna, Utah, and Loma Linda, California, (1974-76)

Data Set		Number of Variables	Number of Observations	Number of Clusters
Angwin	Hour 01	10	48	1
	Hour 05	10	48	2
	Hour 09	10	48	1
	Hour 13	10	48	1
	Hour 17	10	48	1
	Hour 21	10	48	1
	Day 1	10	72	3
	Day 2	10	72	2
	Day 3	10	72	3
	Day 4	10	72	3
	Week 7	10	24	1
	Week 20	10	24	1
	Week 33	10	24	1
	Week 46	10	24	1
	Summer	10	112	4
	Winter	10	112	3
	All Days	10	288	5
Magna	Hour 02	8	48	2
	Hour 05	8	48	1
	Hour 08	8	48	2
	Hour 11	8	48	3
	Hour 14	8	48	2
	Hour 17	8	48	1
	Hour 20	8	48	2
	Hour 23	8	48	2
	Day 1	8	96	1
	Day 2	8	96	2
	Day 3	8	96	3
	Day 4	8	96	2
	Week 7	8	32	1
	Week 20	8	32	2
	Week 33	8	32	1
	Week 46	8	32	1
	Summer	8	112	2
	Winter	8	112	3
	All Days	8	384	4

(continued)

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Table 6 (continued)

Data Set		Number of Variables	Number of Observations	Number of Clusters
Loma Linda	Hour 01	11	48	2
	Hour 05	11	48	2
	Hour 09	11	48	2
	Hour 13	11	48	2
	Hour 17	11	48	2
	Hour 21	11	48	1
	Day 1	11	72	2
	Day 2	11	72	3
	Day 3	11	72	1
	Day 4	11	72	2
	Week 7	11	24	1
	Week 20	11	24	2
	Week 33	11	24	2
	Week 46	11	24	1
	Summer	11	112	2
	Winter	11	112	4
	All Days	11	288	3

EPA-287 (Cin.)
(4-76)

produces the greatest number of clusters for days of the week and for the seasons and year. The inference is made that the days of the week, the seasons, and therefore the year show considerable difference between the clean and dirty situations. At the other two stations that have somewhat less clean situations, the differences between clusters is not as marked.

Magna, Utah, and Loma Linda, California, show an average of two clusters in each hour of the day and for the day of the week examined. By week, these locations show one to two clusters, by season two to three clusters, and by year four to three clusters, respectively.

5.4 MAGNA, UTAH DATA DISCUSSION

The output in terms of tabulations for all three of these CHAMP stations is so great that complete detailed discussion must be limited because of time and space restrictions. These tabulations will be studied in more detail. Also, the number of observations is small in each case, resulting from the need to develop and test the clustering program over as broad a range of situations as possible, yet keeping the computing costs to a minimum.

Only one station was selected for discussion of the output of the clustering program using its pollutant and meteorological data. This station is the Magna, Utah, CHAMP station. Table 7 provides the three cluster output of the NORMIX computer processing of the Magna, Utah, data. The variates are NO, NO_x, O_3, total sulfur, west-east wind component, south-north wind component, temperature, and dew point. All of the pollutant data were transformed by use of logarithms. Where negative variate values were present, a number large enough was added to each value of variate to ensure positiveness before the logarithmic transformation. In this case, the least variable was made equal to 2. Other constants can be selected by the user. There is an option in the program for this selection. Here, 2 served well. The amount needed to be added for each of the four pollutants were respectively 2.0088, 2.0077, 2.0009, and 2.0079. In the case of the mean value of the transformed NO variables in ppm, 0.7071, the anti-ln of 0.7071 is 2.0281. Under the assumption of normality of the logarithms the mean of the logarithm is the mode, median, and mean of the distribution. With back transformation for the above, i.e., the anti-ln, the obtained value 2.0281 is the median value, not the mean value, of the original distribution. When 2.0088 is subtracted from 2.0281, the value of 0.0193 is obtained for the median of the entire sample of variate 1, NO.

Table 7. Three cluster output of NORMIX computer processing of Magna, Utah, day 3, pollutant and meteorological data

```
                                                  PROGRAM NORMIX
                                    WOLFE NORMAL MIXTURE ANALYSIS PROCEDURE(1979 REVISION)

NORMIX PROGRAM WITH MAGNA DATA, DAY 3

      SAMPLE SIZE =          96
      NUMBER OF VARIATES =    8
      NUMBER OF TYPES =       3

                    ITERATION NUMBER   8

   LOG LIKELIHOOD OF  3 TYPES IN THIS SAMPLE =      .98918731+003
                                                  CHARACTERISTICS OF THE WHOLE SAMPLE

     NO          NOX          O3          TS        W COMP      S COMP          T          DP
    PPM          PPM         PPM         PPM          KPH         KPH       DEG C       DEG C
                                                                 MEANS
     1            2           3           4            5           6           7           8
  .7071        .7141       .7070       .7368       -.4513      2.2223     10.7521     -3.6479

                                                            STANDARD DEVIATIONS
     1            2           3           4            5           6           7           8
  .0202        .0266       .0065       .0545       4.6796      9.7852      9.2908      3.5953

                                                                CORRELATIONS
     1            2           3           4            5           6           7           8
 1.0000        .9720      -.6161       .4429       -.0373       .0432      -.4998      -.1393
  .9720       1.0000      -.7145       .4797       -.0539       .0488      -.5766      -.1551
 -.6161       -.7145      1.0000      -.5343        .2342      -.1211       .7780       .2094
  .4429        .4797      -.5343      1.0000       -.0573      -.0337      -.4779      -.1789
 -.0373       -.0539       .2342      -.0573       1.0000      -.5484       .0818       .2510
  .0432        .0488      -.1211      -.0337       -.5484      1.0000       .0548      -.3540
 -.4998       -.5766       .7780      -.4779        .0818       .0548      1.0000       .2806
 -.1393       -.1551       .2094      -.1789        .2510      -.3540       .2806      1.0000

               CHARACTERISTICS OF TYPE    1

           THE PROPORTION OF THE POPULATION FROM THIS TYPE=  .354

     NO          NOX          O3          TS        W COMP      S COMP          T          DP
    PPM          PPM         PPM         PPM          KPH         KPH       DEG C       DEG C
                                                                 MEANS
     1            2           3           4            5           6           7           8
  .7230        .7390       .6978       .7766      -1.6952      3.4025       .4724     -5.5953

                                                            STANDARD DEVIATIONS
     1            2           3           4            5           6           7           8
  .0274        .0317       .0058       .0673       3.4076      4.1199      3.4960      3.3559

                                                                CORRELATIONS
     1            2           3           4            5           6           7           8
 1.0000        .9768      -.4195       .1678        .1272       .0396      -.0482       .1356
  .9768       1.0000      -.4732       .1487        .1523       .0473       .0255       .2303
 -.4195       -.4732      1.0000      -.2589        .1572      -.2832       .3097      -.0207
  .1878        .1487      -.2589      1.0000       -.1004      -.0476      -.0768      -.0894
  .1272        .1523       .1572      -.1004       1.0000      -.5833       .0604       .2631
  .0396        .0473      -.2832      -.0476       -.5833      1.0000       .0063       .1014
 -.0482        .0255       .3097      -.0766        .0604       .0063      1.0000       .7937
  .1356        .2303      -.0207      -.0694        .2631       .1014       .7937      1.0000
```

(continued)

Table 7 (continued)

```
                    CHARACTERISTICS OF TYPE   2

            THE PROPORTION OF THE POPULATION FROM THIS TYPE=  .333

     NO          NOX          O3          TS       W COMP      S COMP        T          DP
    PPM          PPM         PPM         PPM         KPH         KPH      DEG C      DEG C
                                                                MEANS
     1            2           3           4           5           6          7           8
   .6986        .7009       .7132       .7240      3.1261     -8.3204    14.8538    -1.3122

                                                          STANDARD DEVIATIONS
     1            2           3           4           5           6          7           8
   .0021        .0029       .0045       .0346      4.4217      5.5570     5.5933     2.9459

                                                             CORRELATIONS
     1            2           3           4           5           6          7           8
  1.0000        .7217      -.3251       .4652       .3472      -.0605     -.3959      .6380
   .7217       1.0000      -.2914       .6201       .1024       .1709     -.1901      .3358
  -.3251       -.2914      1.0000      -.1993      -.0891       .0766      .6210     -.4192
   .4652        .6201      -.1993      1.0000      -.0449       .4624     -.0026      .1620
   .3472        .1024      -.0891      -.0449      1.0000      -.2701     -.1622      .0539
  -.0605        .1709       .0766       .4624      -.2701      1.0000      .1224     -.2007
  -.3959       -.1901       .6210      -.0026      -.1622       .1224     1.0000     -.4065
   .6380        .3358      -.4192       .1620       .0539      -.2007     -.4065     1.0000

                    CHARACTERISTICS OF TYPE   3

            THE PROPORTION OF THE POPULATION FROM THIS TYPE=  .313

     NO          NOX          O3          TS       W COMP      S COMP        T          DP
    PPM          PPM         PPM         PPM         KPH         KPH      DEG C      DEG C
                                                                MEANS
     1            2           3           4           5           6          7           8
   .6963        .7002       .7108       .7053     -2.6560     12.1212    18.0113    -3.9537

                                                          STANDARD DEVIATIONS
     1            2           3           4           5           6          7           8
   .0036        .0064       .0046       .0115      3.6560      5.9998     6.2153     3.0247

                                                             CORRELATIONS
     1            2           3           4           5           6          7           8
  1.0000        .8374      -.3920       .6050       .3950      -.3067     -.1870      .2724
   .8374       1.0000      -.5435       .8469       .4273      -.3566     -.1960      .2840
  -.3920       -.5435      1.0000      -.5014       .0476       .6293      .3931     -.5575
   .6050        .8469      -.5014      1.0000       .3704      -.3672     -.0876      .0464
   .3950        .4273       .0476       .3704      1.0000       .0739      .1134     -.2557
  -.3067       -.3566       .6293      -.3672       .0739      1.0000     -.0072     -.3433
  -.1870       -.1960       .3931      -.0876       .1134      -.0072     1.0000     -.2240
   .2724        .2840      -.5575       .0464      -.2557      -.3433     -.2240     1.0000
```

Magna, Utah day 3 data statistical characteristics provided in Table 7 established a background against which the cluster characteristics may be compared. The pollutant data are doubly transformed, first by adding a number to each set to assure positiveness and then by taking logarithms. The wind data are transformed from polar to cartesian coordinates. Wind components from the south and/or west are positive. Wind components from the north and/or east are negative. Station to station comparison could be made only by processing mixtures of several stations' data or by comparison with previous output. In the latter case, the numbers added to insure positiveness, or as in some cases, where there is no need to do so, may be different. Complete backtransformation would be required to permit precise comparison. Comparison would then be between medians and probability levels and not means and standard deviations. Table 8 has been prepared to illustrate the problem where the comparison studies are not made in the same run of the data or where adequate back transformation has not yet been made. Table 8 shows comparative output for (1) Angwin, California; (2) Loma Linda, California, and (3) Magna, Utah for the means and standard deviations of the three variates (1) (NO), (2) (NO_x) and (3) (O_3). With the exception of the Loma Linda (NO_x) all data were scaled upwards to a minimum value of two to insure positiveness. Logarithms were then used to transform all data.

Table 8. Intercomparison of selected pollutant data through the year (1) (NO), (2) (NO_x) and (3) (O_3) among (1) Angwin, California, (2) Loma Linda, California and (3) Magna, Utah. The values are given in logarithmically transformed data orginally given in ppm.

Transformed variates	(NO)		(NO_x)		(O_3)		No. of Data
Locations	mean	s.d.	mean	s.d.	mean	s.d.	
Angwin, CA	0.6981	0.0017	0.7002	0.0045	0.7127	0.0087	288
Loma Linda, CA	0.7036	0.0130	-3.0442	0.9576	0.7081	0.0198	112
Magna, UT	0.7055	0.0164	0.7123	0.0216	0.7081	0.0090	324

9-1/8
BOTTOM OF IMAGE AREA, OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

The information in Table 8 reveals quite clearly that Angwin, California, has probably the lowest oxides of nitrogen and highest ozone pollutant level and the least variability of these three variates of the three stations. This is one of the reasons that the Angwin site was chosen. The (NO_x) value of -3.0442 with the standard deviation of .9576 reflects the non addition of a number to insure a minimum value of 2 before logarithms were taken. This non-transformation was used in order to demonstrate that, insofar as the NORMIX program is concerned, negative numbers can be used. The assumption of normality may or may not be valid. For example, the additive number for the (NO_x) at Magna, Utah, was 2.0077. The anti-logarithm of 0.7123 is 2.0385. The difference is 0.0308 ppm. The anti-logarithm of -3.0442 is 0.0467 ppm at Loma Linda, California. These are median values, not necessarily mean values of the original data.

Examination of the Magna, Utah, data shown in Table 7, day 3 only reveals interesting aspects of the relationships of the pollutant and meteorological distributions. A few are now noted.

The variates' averages and standard deviation values of the transformed pollutant variables are of (1) the oxide of nitrogen (NO), 0.7171 and 0.0202 ppm, (2) the oxides of nitrogen (NO_x) 0.7141 and 0.0266 ppm, (3) ozone (O_3), 0.7070 and 0.0085 ppm, and (4) total sulfur (TS), 0.7368 and 0.0545 ppm. The correlations of (NO), (NO_x) and (TS) with (O_3) are all negative with respective values of -0.6181, -0.7145 and -0.5343. This inverse relationship is noted with few exceptions throughout most summary output.

Equally interesting is the positive relationship of ozone (O_3) with temperature and dew point in the total ensemble changing to negative for the dew points within the clusters. This feature is lost when the entire set of all data available are used for the selected hours of the day, the days of the week and the weeks of the year. The positive relationship with the dew points remains. See Table 17 of Volume III of this report. Information from Tables 13 and 15 from Volume III was used also to construct Table 8 of this report. Peculiar is the fact in Table 7 that the variability of the (NO), (NO_x) and (TS) is greater in Cluster 1 than in the total ensemble while in Clusters 3 and 2, in that order, the variability is less. This feature also holds for Cluster 4 of the annual set of data shown in Table 17 of Volume III of this report. The variability of the remaining elements is less

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
EPA-287 (Cin.)
(4-76)

in all clusters than in the total ensembles. As the mean and variance, especially in Cluster 1, seem to be positively correlated, one inference that can be drawn is that the logarithmic transformation was insufficient. An alternative inference, equally speculative, is that further processing might allow Cluster 1 to break down into two smaller clusters.

Figure 14 developed from Table 7 output data shows the 0.50 probability ellipses and the associated pollutant means, standard deviations, proportions and the numbers of observations. The classified data overplot may be examined to help assess the clustering efficacy. It must be remembered that the pollutant variables have been logarithmically transformed and all numbers here refer to transformed values.

Figure 15 presents some of the same information in profile form, variate by variate, against a standardized three sigma background. The relationships among the variates can be further assessed with this different presentation. As mentioned previously, the order of the variates may be changed. This is done here. The inverse relationship of ozone with the oxides of nitrogen and sulfur is seen clearly.

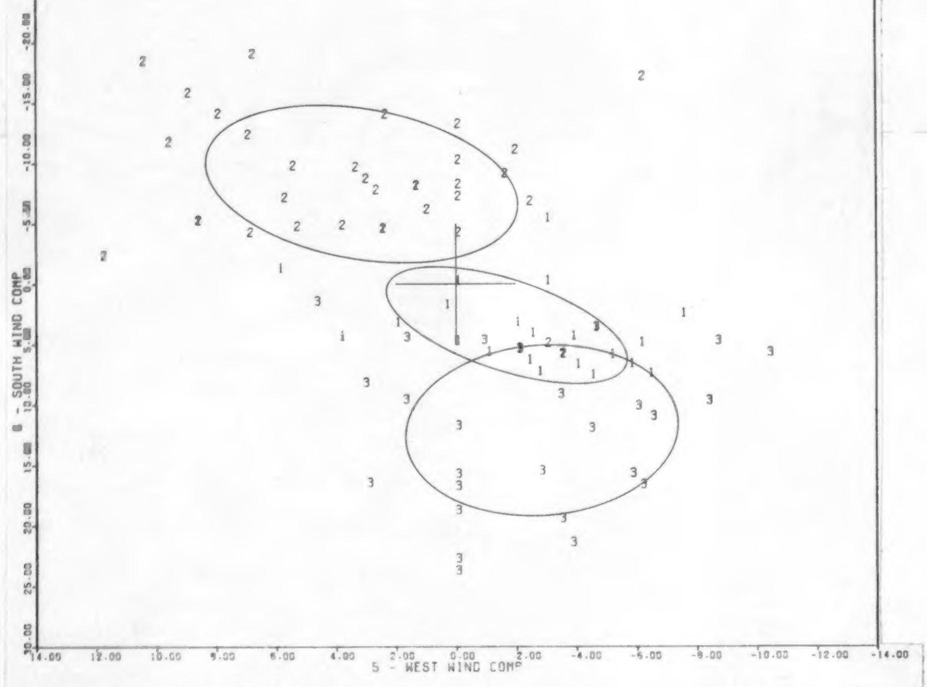
There are limitations to all presentation forms because of the many relationships among variates. With p variates there are p(p-1)/2 paired relationships for the total ensemble and for each cluster. Then there are the triplet relationships and the higher order marginal relationships which might be presented if ways to present them are developed. Here in Figure 14 the bivariate plot of the wind components is only one of the 36 similar paired plots that can be made with 8 variates. Some idea of correlation between variates, as shown in Table 7, can be deduced from the ellipse orientations as shown in Figure 14 and by the trend changes which may be noted in Figure 15.

Weak southeast winds with a vector mean of 4 kph are associated with lower than average ozone readings and higher than average readings of the oxides of nitrogen and sulfur. Strong winds from the northwest, 9 kph, and from the south-southeast, 12 kph, show the inverse relationship of higher ozone and lower oxides of nitrogen and sulfur. Clusters 1 and 3, with southeast and south-southeast winds, respectively, show the greater temperature difference between the means, namely 17.54° C.

Volume III of this report provides more examples of cluster analysis of the CHAMPS data and includes more detailed comments.

MAGNA, DAY 3 N=96

VARIABLE	CLUSTER 1	CLUSTER 2	CLUSTER 3
	P= .354	P= .333	P= .313
1 - NO			
MEAN	0.72	0.70	0.70
STANDARD DEVIATION	0.03	0.00	0.00
2 - NOX			
MEAN	0.74	0.70	0.70
STANDARD DEVIATION	0.03	0.00	0.01
3 - OZONE			
MEAN	0.70	0.71	0.71
STANDARD DEVIATION	0.01	0.00	0.00
4 - TS			
MEAN	0.78	0.72	0.71
STANDARD DEVIATION	0.07	0.03	0.01
5 - WEST WIND COMP			
MEAN	-1.70	3.13	-2.86
STANDARD DEVIATION	3.41	4.42	3.86
6 - SOUTH WIND COMP			
MEAN	3.40	-8.32	12.12
STANDARD DEVIATION	4.12	5.56	6.00
7 - TEMPERATURE			
MEAN	0.47	14.85	18.01
STANDARD DEVIATION	3.49	5.59	6.22
8 - DEW POINT			
MEAN	-5.60	-1.31	-3.93
STANDARD DEVIATION	3.36	2.95	3.02



WEST VS SOUTH WIND - PROBABILITY LEVEL = 0.50

Figure 14. Magna, Utah, day 3, 0.50 probability ellipses of the west-east and south-north wind components for three cluster types. Winds from the west and south are positive.

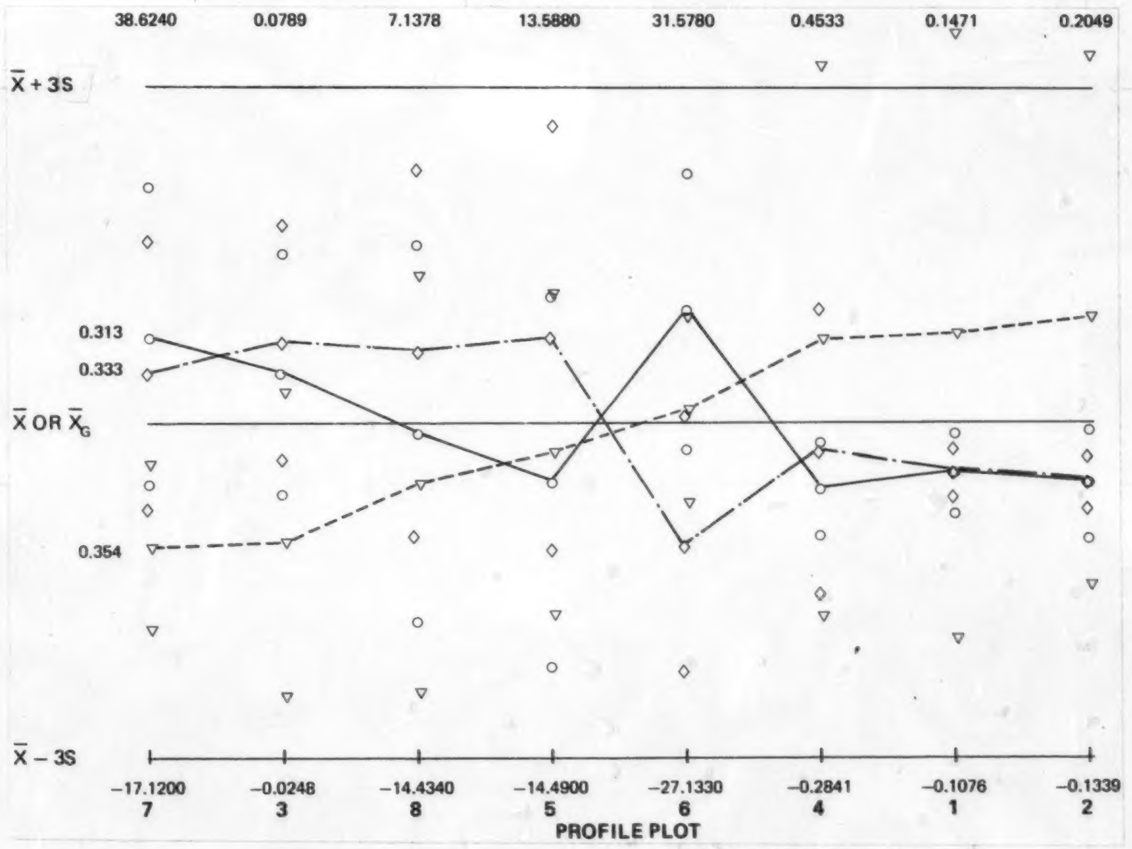


Figure 15. Magna, Utah, day 3, profile plots of pollutants and meteorological variates against a three sigma background.

6.0 RESULTS AND CONCLUSIONS

Reference is made to the three-fold objective:

1. clustering
2. validation, and
3. classification of environmental data.

A computer program, tested and calibrated, has been placed in the computer library of EPA-NCC and is available to fulfill the three-fold objective. This program is a modified version of the Wolfe NORMIX computer program. A number of options have been added to this version and made available for the user to employ prior to and after the central NORMIX core.

Clustering is attained in the first portion of the central program. Those singleton data or small clusters that are the last to merge with the larger clusters are logical candidates for examination as potential "outliers". "Outlier" treatment was discussed. Data should be processed through a pre-edit routine to remove impossible data values and to make certain that all elements of a multivariate (multi-element) vector are present. The initial clusters enter the secondary portion of the program. Statistical treatment of these clusters results in the establishment of their characteristics to permit classification. A number of options have been added to permit the user to examine the distributions in more detail. The increased output permits graph or plot comparison of data in profiles and/or in selected bivariate representations.

The computer program can be used to validate pollutant and meteorological as well as any other type of data. Three CHAMP station data files were used for this report. Significant clustering was achieved.

This program can be used to examine and to validate large blocks of data. However, it is suggested that the processing be done in small batches to keep computing costs to a minimum. If the classification portion is used for study and validation, preliminary estimates of cluster statistics taken from previous batch runs will reduce the overall costs.

The computer program output when processing environmental data can serve to indicate pollutant sources or episodes.

BOTTOM OF IMAGE AREA, OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

7.0 REFERENCES

Aitchison, J., and J. A. C. Brown. 1976. The Lognormal Distribution. Cambridge University Press, New York, New York. 176 pp.

Anderson, Edgar. 1935. The Irises of the Gaspé Peninsula. Bull. Amer. Iris Soc. 59:2-5.

Barnard, M. A. 1936. The Secular Variations of Skull Characters in Four Series of Egyptian Skulls. Ann. Eugen. 6:352-371.

Barnett, V., and T. Lewis. 1979. Outliers in Statistical Data. John Wiley and Sons, New York, New York. 365 pp.

Barr, A. J., J. H. Goodnight, J. P. Sale, and J. T. Helwig. 1976. A User's Guide to SAS. Spanks Press.

Bartlett, M. S. 1937. Properties of Sufficiency and Statistical Tests. Proc. R. Soc. A. 160:268-282.

Binder, D. A. 1978. Bayesian Cluster Analysis. Biometrika 65:31-38.

Bose, R. C. and S. N. Roy. 1938. The Distribution of the Studentized D^2 Statistic. Sankhya, 4:I:19-38.

Brooks, C. E. P., C. S. Durst, and N. Carruthers. 1946. Upper Winds Over the World. Part 1. The Frequency Distribution of Winds at a Point in the Free Air. Q.J.R. Meteor. Soc. 72:55-73.

Chauvenet, W. 1863. Method of Least Squares, Appendix to Manual of Spherical and Practical Astronomy. J. Lippincott and Co., Philadelphia, Pennsylvania. 2:469-566, Tables 593-599, 5th ed., Dover, New York, New York. (Out-of-print)

Cormack, R. M. 1971. A Review of Classification. J. R. Statist. Soc. Series A 134:321-367.

Creason, J. P. 1978. The Theory and Application of a General Iterative Maximum Likelihood Procedure to Randomly Censored Univariate and Bivariate Linear Models. Inst. of Statist. Mimeo Series, No. 1173, Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina. 137 pp.

Crutcher, H. L. 1957. On the Standard Vector Deviation Wind Rose. J. Meteor. 14:1:28-33.

Crutcher. H. L. 1960. Statistical Grouping of Climates and the Statistical Discrimination among Climatic Groups. Ph.D. Dissertation, New York University, New York, New York. Xerox University Microfilms, Ann Arbor, Michigan. 462 pp.

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER

Crutcher, H. L. 1979. Cluster Analysis as a Data Validation Technique. In: Data Validation Conference Proceeding, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina. EPA-600/9-79-042. 72-80.

Crutcher, H. L., and L. W. Falls. 1976. Multivariate Normality. U.S. National Aeronautics and Space Administration, Washington, D.C. NASA TN D-8226. 273 pp.

Crutcher, H. L., and R. L. Joiner. 1977. Separation of Mixed Data Sets into Homogeneous Sets. U.S. Department of Commerce, NOAA Technical Report EDS-19. 167 pp.

Crutcher, H. L., and R. L. Joiner. 1978. Another Look at the Upper Winds of the Tropics. J. Meteor. 16:5:462-476.

Dixon, W. J. 1950. Analysis of Extreme Values. Ann. Math. Statist. XXI:488-505.

Duran, B. S., and P. L. Odell. 1974. Cluster Analysis, A Survey. Lecture Notes in Economics and Mathematical Systems, Econometrics. Springer Verlag, New York, New York. 137 pp.

Everitt, B. S. 1974. Cluster Analysis. Heinemann, London, England.

Everitt, B. S. 1979. Unresolved Problems in Cluster Analysis. Biometrics. 35:169-181.

Fairbairn, B., A. C. Nelson, Jr., and M. Symons. 1979. Investigation and Report of Cluster Analysis. Prepared for U.S. Environmental Protection Agency, Research Triangle Park, North Carolina. PEDCo Environmental, Inc., Contract No. 68-02-2722, Task 11. 30 pp.

Ferguson, T. S. 1961. On the Rejection of Outliers. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. 1:253-289.

Ferguson, T. S. 1961. Rules for Rejection of Outliers. Rev. Inst. Int. de Statist. 29:29-43.

Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugen. VII:11:179-188.

Fitz-Simons, T., and D. M. Holland. 1979. The Maximum Likelihood Approach to Probabilistic Modeling of Air Quality Data. U.S. Environmental Agency, Research Triangle Park, North Carolina. EPA-600/4-79-044: 1-34.

Fix, E., and J. L. Hodges, Jr. 1951. Discriminatory Analysis Nonparametric Discriminant Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas, Report No. 4. 121 pp.

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Friedman, H. P., and J. Rubin, 1967. On Some Invariant Criteria for Grouping Data. J. Am. Statist. Assoc. 62:1159-1178.

Gosset, W. S. (Student). 1925. New Tables for Testing the Significance of Observations. Metron. 5:3:105-120.

Green, P. E. with J. D. Carroll. 1976. Mathematical Tools for Applied Multivariate Analysis. Academic Press, New York, New York. 376 pp.

Grubbs, F. E. 1950. Sample Criteria for Testing Outlying Observations. Ann. Math. Statist. 21:27-58.

Grubbs, F. E. 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics, 11:1-21.

Hald, A. 1952. Statistical Theory with Engineering Applications. John Wiley and Sons, New York, New York. 783 pp.

Harter, H. L. 1974-76. The Method of Least Squares and Some Alternatives. I-VI, Rev. Int. Inst. de Statist. 42:147-174:Part I, 42:235-264:Part II, 43:1-44:Part III, 43:125-190:Part IV, 43:269-278:Part V, 44:113-159:Part VI.

Hartigan, J. A. 1975. Clustering Algorithms. John Wiley and Sons, New York, New York. 351 pp.

Hotelling, H. 1936. Relations Between Two Sets of Variates. Biometrika, 28:321-377.

Hotelling, H. 1954. Multivariate Analysis. In: Statistics and Mathematics in Biology. eds. O. Kempthorne et al. Iowa State College Press, Ames, Iowa. 67-80.

International Mathematical and Statistical Laboratories, Inc. 1977. Computer Subroutine Libraries in Mathematics and Statistics. IMSL LIB-0006, Chapters O, S, and U.

Kendall, M. G. 1966. Discrimination and Classification. In: Multivariate Analysis. ed. P. R. Krishnaiah. Academic Press, New York, New York. 165-184.

Klint, W. E. 1979. Screening Checks Used by the National Climatic Center. In: Data Validation Conference Proceedings, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina. EPA-600/9-79-042: 135-198.

Larsen, R. I. 1971. A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards. U.S. Environmental Protection Agency, Office of Air Programs Publication, No. AP-89. 55 pp.

Lee, K. L. 1979. Multivariate Tests for Clusters. J. Am. Statist. Assoc. 74:367;708-714.

BOTTOM OF IMAGE AREA, OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Likeš, J. 1966. Distribution of Dixon's Statistics in the Case of an Exponential Population. Metrika. 11:46-54.

Mage, D. T., and W. R. Ott. 1978. Refinements of the Lognormal Probability Model for Analysis of Aerometric Data. J. Air Poll. Control Assoc. 28:8:796-798.

Mahalanobis, P. C. 1936. On the Generalized Distance in Statistics. Proceedings of the Nat. Inst. of Science of India. II:1:45-55.

Marsaglia, G. K., N. Anathanarayan, and N. Paul. 1972. Random Number Generator Package, "Super-Duper" Uniform, Normal and Random Number Generator. McGill University, School of Computer Science Computing Center, Montreal, Canada.

McRae, D. J. 1973. MIKCA: A FORTRAN IV Iterative K-Means Cluster Analysis Program. CTB/McGraw Hill, Del Monte Research Park, Monterey, California. 1970. Revised by M. J. Symons, October 1973.

Miller, R. G. 1962. Statistical Prediction by Discriminant Analysis, Meteorological Monographs. Am. Meteor. Soc., Boston, Massachusetts. 4:25. 54 pp.

Moore, P. G. 1957. Transformations to Normality Using Fractional Powers of the Variable. J. Am. Statist. Assoc. 52:278:237-246.

Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent. 1975. SPSS, Statistical Package for the Social Sciences. 2nd ed. McGraw-Hill, Inc. 249-275.

Pearson, K. 1894. Contributions to the Mathematical Theory of Evolution. Phil. Trans of the R. Soc. of London, Series A 185:71-110 + 5 plates.

Pearson, K. 1901. On the Lines and Planes of Closest Fit to Systems of Points in Space. Phil. Mag. 6:6:559-572.

Pearson, E. S., and C. Chandra-Sekar. 1936. The Efficiency of Statistical Tools and A Criterion for the Rejection of Outlying Observations. Biometrika 28:308-320.

Rao, C. 1952. Advanced Statistical Methods in Biometric Research. John Wiley and Sons, Inc., New York, New York, 1-390.

Rhodes, R. C. and S. Hochheiser. 1979. Data Validation Conference, ORD-USEPA. 600/9-79-042. 316 pp.

Suñé, J. M., D. Pizer, and G. Sharp. 1978. Air Pollutants in Selected Salt Lake Basin Communities, 1963-1976. Exposure Assessment Branch, Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, unpublished.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

Thompson, W. R. 1935. On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation. Ann. Math. Statist. 6:214-219.

U.S. Environmental Protection Agency. 1979. Data Validation Conference Proceedings. Research Triangle Park, North Carolina. EPA-600/9-79-042. 316 pp.

Ward, J. H., Jr. 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Statis. Assoc. 58:236-244.

Ward, J. H., Jr., K. Hall, and J. Buchhorn. 1967. PERSUB Reference Manual, PRL-TR-67-3 (II). Personnel Research Laboratory, Lackland Air Force Base, Texas. 60 pp.

Wishart, D. 1978. Clustan User Manual. 3rd ed. Program Unit, Edinburgh University, Edinburgh, Scotland.

Wolfe, J. H. 1965. A Computer Program for the Maximum Likelihood Analysis of Types. Technical Bulletin 65-15, U.S. Naval Personnel Research Activity, San Diego, California. AD 620 020. 52 pp.

Wolfe, J. H. 1967a. NORMIX Cluster and Pattern Analysis of Normal Mixtures. FOCUS-Library-G7-USNP-NORMIX, FORTRAN COOP-COOP Monitor, U.S. Naval Personnel Research Activity, San Diego, California. 82 pp.

Wolfe, J. H. 1967b. NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixture of Distributions. Research Memorandum SRM-2, U.S. Naval Personnel Research Activity, San Diego, California. AD 656 588. 30 pp.

Wolfe, J. H. 1968. NORMAP: Cluster and Pattern Analysis of Normal Mixtures with Common Covariances. FOCUS-Library-G7-USNP-NORMAP, FORTRAN 63 COOP-COOP MONITOR-1604. U.S. Naval Personnel Research Activity, San Diego, California. 88 pp.

Wolfe, J. H. 1970. Pattern Clustering by Multivariate Analysis. Multivariate Behavioral Research. 5:329-350.

Wolfe, J. H. 1971a. A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. Technical Bulletin STB-72-2. Naval Personnel and Training Research Laboratory, San Diego, California. AD 731 039. 7 pp.

Wolfe, J. H. 1971b. NORMIX 360 Computer Program. Research Memorandum SRM 72-4. Naval Personnel and Training Research Laboratory, San Diego, California. 125 pp.

Wright, T. W. 1884. A Treatise on the Adjustment of Observation by Method of Least Squares. Van Nostrand, New York, New York.

8.0 BIBLIOGRAPHY

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York, New York, pp. 1-359.

Ball, G. H., and D. H. Hall. 1967. A clustering technique for summarizing multivariate data. Behavioral Science. 12:2:153-155.

Blifford, I. H., Jr. and G. O. Meeker. 1967. A factor analysis model of large scale pollution. Atmospheric Environment, Pergamon Press. 1:147-157.

Chernoff, H. 1971. The use of faces to represent points in n-dimensional space graphically. Technical Report No. 71, Department of Statistics, Stanford University, Stanford, California. pp. 1-45.

Cooley, W. W. and P. R. Lohnes. 1962. Multivariate procedures for the behavioral sciences. John Wiley and Sons. New York, New York. pp. 1-211.

Gatz, D. F. 1978. Identification of Aerosol Sources in the St. Louis area using factor analysis. Jour. App. Meteor. 17:600-608.

Harman, N. H. 1967. Modern factor analysis. University of Chicago Press, Chicago, Illinois. pp. 1-474.

Hopske, P. K., E. S. Gladney, G. E. Gordon, W. H. Zoller and A. G. Jones. 1976. The use of multivariate analysis to identify sources of selected elements in the Boston urban aerosol. Atmospheric Environment, Pergamon Press. 10:1015-1025.

Kowalski, B. R. and C. F. Bender. 1972. Pattern recognition. A powerful approach to interpreting chemical data. Jour. Am. Ch. Soc. 94:16:5632-5640.

Ling, R. F. 1973. A probability theory of cluster analysis. Jour. Am. Statist. Assoc. 68:341:159-164.

Larsen, R. I. 1969. A new mathematical model of air pollutant concentration averaging time and frequency. Jour. Air Pollution Control Assoc. 19:1:24-30.

Larsen, R. I. 1973. An air quality data analysis system for interrelating effects, standards, and needed source reductions. Jour. Air Pollution Control Assoc. 23:11:933-940.

Larsen, R. I. 1974. An air quality data analysis system for interrelating effects, standards, and needed source reductions. 24:6:551-550.

Larsen, R. I. 1976. An air quality data analysis system for interrelating effects, standards, and needed source reductions - part 3. Vegetation injury. Jour. Air Pollution Control Assoc. 26:4:324-333.

Larsen, R. I. and W. W. Heck. 1977. An air quality data analysis system for interrelation effects, standards, and needed source reductions - part 4. A three parameter averaging-time model. Jour. Air Pollution Control Assoc. 27:5:454-459.

Larsen, R. I., D. E. Gardner, and D. L. Coffin. 1979. An air quality analysis system for interrelating effects, standards, and needed source reductions - part 5. NO_2 mortality in mice. Jour. Air Pollution Control Assoc. 29:2:133-137.

Neustadter, H. E., J. S. Fordyce, and R. B. King. 1976. Elemental Composition of airborne particulates and source identification: Data analysis techniques. Jour. Air Pollution Control Assoc. 26:11:1079-1089.

Prinz, Bernd and H. Stratman. 1968. The possible use of factor analysis in investigating air quality. Staub-Reinhalt. 28:1:33-39.

Ruff, R. E. 1973. Application of adaptive pattern classification to the derivation of relationships between air quality data. Paper presental at the International Symposium on "Automated Air Quality Monitoring Networks and Data Analysis, Bilthoven. The Netherlands. pp. 1-25.

Sokal, R. R. and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Company, San Francisco, California. pp. 1-359.

Tryon, R. C. and D. E. Bailey. 1970. Cluster analysis. McGraw-Hill Book Company, New York, New York. pp. 1-347.

BEGIN LAST LINE OF TEXT

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS

3/8"

PAGE NUMBER

EPA-287 (Cin.)
(4-76)

TYPING GUIDE SHEET
BEGIN FIRST LINE OF TEXT HERE
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD; BEGIN SECTIONS HERE
6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

PART 2

INVESTIGATION AND REPORT
OF
CLUSTER ANALYSIS

INVESTIGATION AND REPORT
OF
CLUSTER ANALYSIS

6-1/2"
9-1/8"

Prepared by

B. Fairbairn, C. Nelson, and M. Symons
PEDCo Environmental, Inc.
505 South Duke Street
Durham, North Carolina 27701

Contract No. 68-02-2722
Task 11

Project Officer
Raymond C. Rhodes
Environmental Monitoring
Systems Laboratory

Project Manager
Gerald Akland
Environmental Monitoring
Systems Laboratory

Prepared for

U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

April 16, 1979

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

P2

TABLE OF CONTENTS

Page

LIST OF TABLES iii

LIST OF FIGURES iv

1.0 INTRODUCTION 1-1

2.0 DATA DESCRIPTION 2-1

3.0 CLUSTER ANALYSIS 3-1

3.1 Introduction 3-1

3.1.1 Limitations of the Methods 3-1
3.1.2 Application of Methods 3-2
3.1.3 The SAS Cluster Algorithm 3-2
3.1.4 MIKCA Algorithm 3-5
3.1.5 NORMIX Algorithm 3-7

3.2 Analysis of 3-7 PM Daily Average Data 3-7
3.3 Analysis of Four 1-Hour Periods Per Day 3-11
3.4 Applications of Cluster Analysis 3-16

4.0 MULTIPLE REGRESSION ANALYSIS 4-1

5.0 CONCLUSIONS 5-1

6.0 REFERENCES 6-1

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

LIST OF TABLES

Page

2-1 Distribution of Hourly Observations, 3 PM to 7 PM 2-1
3-1 Possible Outlier Observations 3-8
3-2 Cluster Centers of Eight Groups: 3-7 PM Daily Average Data 3-10
3-3 Cluster of Four Groups 3-7 PM Daily Average Data 3-11
3-4 Frequency Tables By Day of Week and Month for MIKCA Results for Four Groups 3-7 PM Daily Average Data 3-12
3-5 Comparison of MIKCA and SAS Cluster Centers for Four Groups: 3-7 PM Daily Average Data 3-13
3-6 Comparison of MIKCA and SAS Cluster Centers for Three Groups: 3-7 PM Daily Average Data 3-13
3-7 Cluster Centers for Eight-Group Analysis for Four 1-Hour Periods Per Day Data 3-15
3-8 Cluster Center for Four-Group Analysis: Criterion to Minimize Determinant W for Four 1-Hour Periods Per Day Data 3-15
3-9 Cluster Center for Four-Group Analysis: Criterion to Minimize Bayes Factor 1 of Determinant W for Four 1-Hour Periods Per Day Data 3-16

6-1/2"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

LIST OF FIGURES

		Page
1-1	Los Angeles Catalyst Study Site	1-2
3-1	Tree Diagram	3-3
3-2	Cluster Map	3-4
3-3	General Flow-Density Relationship	3-9

6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

1.0 INTRODUCTION

This project was initiated to investigate cluster analysis as a multivariate tool for the preliminary modeling of air pollution data. The data for this investigation originated with the Los Angeles Catalyst Study (LACS).[1] The LACS was initiated to investigate the impact of catalyst-equipped cars on the ambient air. To do this, it was necessary to gather data before and after 1975, when catalytic converters were installed on cars. Thus the data base was begun in June 1974, and the data collection is continuing.

Two sites on each side of the San Diego Freeway near its intersection with Sunset Boulevard (Figure 1-1) were chosen for LACS because they were expected to have the least interference from pollutants from the many citywide industries and the airport. At the sites, prevailing wind blows from the Pacific Ocean directly across the highway. Thus, the upwind sites on the west side of the freeway were expected to monitor the background pollutant levels, while the sites immediately east of the freeway monitor the background levels plus any others from the highway when the wind is from the ocean. A breeze towards the ocean could have different background pollutant levels than a breeze blowing inland, but as long as data are collected on both sides of the freeway, the traffic effect can be discerned.

Many pollutants and various meteorological factors were monitored on both sides of the highway. Data on levels of ozone (O_3), nitrogen dioxide (NO_2), nitrogen oxide (NO), lead (Pb), suspended particles (SP), carbon monoxide (CO), and sulfur dioxide (SO_2) were recorded. Meteorological data included temperature, windspeed, and wind direction. The San Diego Freeway has counter loops under all 10 lanes (5 northbound, 5 southbound) which record by lane the traffic counts in speed increments of 10 mph.

Much work has already been done with the data base, and new studies are being conducted since the data are continuously being updated.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-1

P6

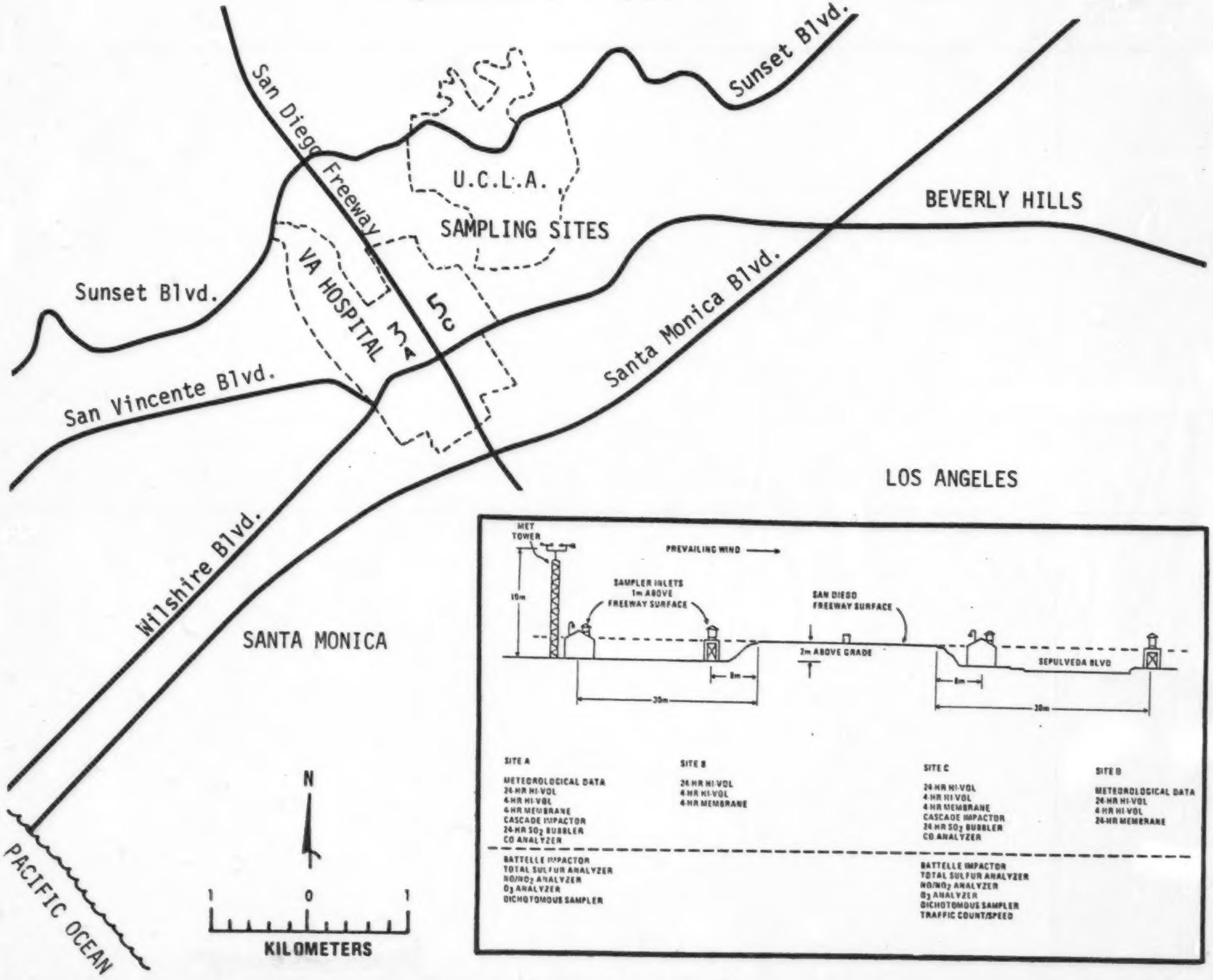

Sunset Blvd.
San Diego Freeway
U.C.L.A.
SAMPLING SITES
BEVERLY HILLS
Sunset Blvd.
VA HOSPITAL
3 A
5 C
San Vincente Blvd.
Santa Monica Blvd.
Wilshire Blvd.
LOS ANGELES
SANTA MONICA
PACIFIC OCEAN
N
1
0
1
KILOMETERS
MET TOWER
PREVAILING WIND
10m
SAMPLER INLETS 1m ABOVE FREEWAY SURFACE
SAN DIEGO FREEWAY SURFACE
2m ABOVE GRADE
SEPULVEDA BLVD
8m
30m
8m
30m
SITE A
METEOROLOGICAL DATA
24-HR HI-VOL
4-HR HI-VOL
4-HR MEMBRANE
CASCADE IMPACTOR
24-HR SO2 BUBBLER
CO ANALYZER
BATTELLE IMPACTOR
TOTAL SULFUR ANALYZER
NO/NO2 ANALYZER
O3 ANALYZER
DICHOTOMOUS SAMPLER
SITE B
24-HR HI-VOL
4-HR HI-VOL
4-HR MEMBRANE
SITE C
24-HR HI-VOL
4-HR HI-VOL
4-HR MEMBRANE
CASCADE IMPACTOR
24-HR SO2 BUBBLER
CO ANALYZER
BATTELLE IMPACTOR
TOTAL SULFUR ANALYZER
NO/NO2 ANALYZER
O3 ANALYZER
DICHOTOMOUS SAMPLER
TRAFFIC COUNT/SPEED
SITE D
METEOROLOGICAL DATA
24-HR HI-VOL
4-HR HI-VOL
24-HR MEMBRANE

Figure 1-1. Los Angeles Catalyst Study Site.

PAGE NUMBER
EPA-287 (Cin.)
(4-76)

2.0 DATA DESCRIPTION

The LACS data were originally on two tapes: one had meteorological and pollutant data, and the other had traffic count data. A few problems with the tape containing the meteorological and pollutant data caused a considerable time delay before data manipulation could begin. Initially, there were 25,133 daily observations; however, to obtain complete data sets, these were pared by eliminating data from sites 4, 6, 7, and 8; by excluding data from only the period 9/14/77 to 6/1/78; by deleting duplicated data; and by including only sites 3 and 5. Table 2-1 shows the number of hourly observations by variable and site.

6-1/2"

The next step was to take daily averages between the hours of 3 PM and 7 PM for each variable. The results were seven data files--one for each of the following: suspended particles, lead, carbon monoxide, windspeed, wind direction, temperature, and traffic counts.

TABLE 2-1. DISTRIBUTION OF HOURLY OBSERVATIONS, 3 PM to 7 PM

Variable	Number of Observations Site 3	Site 5
Suspended particles	132	135
Lead	125	135
Carbon monoxide	384	384
Windspeed	451	0
Wind direction	449	0
Temperature	436	157
Total	1977	811

One data set was made of the traffic data for the 1977-78 period named above. Using site 5 for the pollutant data and site 3 for meteorological data, the data were combined from both sites to get a more complete final data set for analysis. Thus seven files were merged by dates to make a single observation of several variables for each day. Days were deleted when one or more of the variables were missing on that date. For the 3-7 PM daily averages over the period 9/14/77 through 6/1/78, there were 63 complete data;

BOTTOM OF IMAGE AREA OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8
PAGE NUMBER
(4-76)

these daily data made up one of the two data sets used in the final cluster and the multivariate analyses.

Hourly data were also of interest, so the same data manipulation was applied to the hourly observations for each variable. However, lead and suspended particles had to be dropped from this data set because the data were coded only once or twice daily. After the final merging of the data and deletion of observations with missing values, the data contained 3031 hourly observations. This total represents approximately half of the possible observations (study period of 261 days × 24 hours = 6264 hourly observations). Of those 3031, there were 97 days with more than 20 observations and 32 days with 24 observations. Because most clustering techniques cannot handle over 1000 observations, another file was constructed from four 1-hour periods for each day. The hours chosen were 3 AM, 8 AM, 2 PM, and 5 PM. The resulting file contained 523 observations.

9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

3.0 CLUSTER ANALYSIS

3.1 INTRODUCTION

Cluster analysis is a multivariate statistical method of preliminary analysis used to reveal any inherent structure of the data. The basic steps of the analysis are to divide the data into groups which optimize a criterion and to examine these groups of observations for any similarities that help in understanding any interrelationships of the variables observed.

3.1.1 Limitations of the Methods

There are several limitations in cluster analyses. There is no guarantee that cluster analyses will necessarily find the desired natural groups that will reveal underlying distributions of variables. Under some but not all circumstances, grouping the variables will reveal which ones are in some way related.

Another limitation can be the special structure that hierarchical clustering imposes on the data. Hierarchical clustering divides N items of data into n groups (where $1 \leq n \leq N$). In going from n to n + 1 groups, n - 1 of the groups are identical, and the extra remaining group (in this case of n groups) is split by a criterion function to form (n - 1) + 2 or n + 1 groups.

In theory, the hierarchical algorithm can be achieved by either the divisive or the agglomerative method. The divisive method first considers all the data as one single group; then divides the data into two groups; and finally subdivides one of the two groups to get a third group. This process continues until there are N groups--i.e., each datum is a group. The agglomerative method proceeds in the opposite direction from the divisive method. Initially, it considers each datum as a group, and then it combines two groups (two of the N data elements) that are by some criterion the "closest."

The divisive algorithm is, in general, extremely difficult computationally. On the first pass through the data there are 2^N-1 possible two groups for which some criterion value must be computed for each combination before deciding on the best possible split of the data into two groups. This number, 2^N-1, becomes large very fast; in fact, for $N > 16$, a complete enumeration is

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

not practical.[2] In contrast, the agglomerative algorithm requires considerably fewer computations.

The final groups of data from using either of the nested hierarchical algorithms will result in a "tree" diagram, or a dendogram. The divisive algorithm starts at the top of the tree and descends, splitting and branching into single data items at the bottom of the tree. The agglomerative algorithm starts at the bottom and joins data together until all are joined into one group at the top.

Consider the following example. Each of eight data elements, which could consist of any number of variables, is represented by one of the letters A-H. Suppose a hierarchical algorithm resulted in the tree diagram in Figure 3-1. If the divisive method had been imposed, the splitting would have begun at the top at the dashed line and would have separated H and D into one group and F, E, A, G, B, and C into a second group. If the agglomerative method had been used, the first joining would have been between G and B (resulting in seven groups), and the second by the joining of F and E (leaving six groups), and on up the tree until there would be only one group.

The main drawback to the agglomerative method is that a link between two groups, once made, can never be broken. Finding the natural grouping of the data thus becomes restricted.

3.1.2 Application of the Methods

Two computer programs were used to examine the data: a SAS procedure and MIKCA (an iterative K-means cluster analysis program). A third program, NORMIX, will be discussed, but it was not available for analysis of either data base.

3.1.3 The SAS Cluster Algorithm

SAS, Statistical Analysis System, is a computer software package that incorporates many statistical analysis algorithms.[3] The algorithm used on the project data base (called CLUSTER) is a hierarchical clustering routine which uses the agglomerative approach by considering each datum as an individual group and by continuing the grouping until all the data are in one group. This is a complete linkage algorithm.

The clustering map (Figure 3-2) is comparable to the tree diagram illustrated earlier. The identifier associated with each datum (data item) is printed across the top of the page, and below each is an asterisk (*) to signify that at that point each datum is a separate group. Beside the line of

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
(4-76)

BEGIN
FIRST
LINE OF
TEXT
HERE

TYPING GUIDE SHEET

CENTER
OF PAGE

TOP OF
IMAGE
AREA

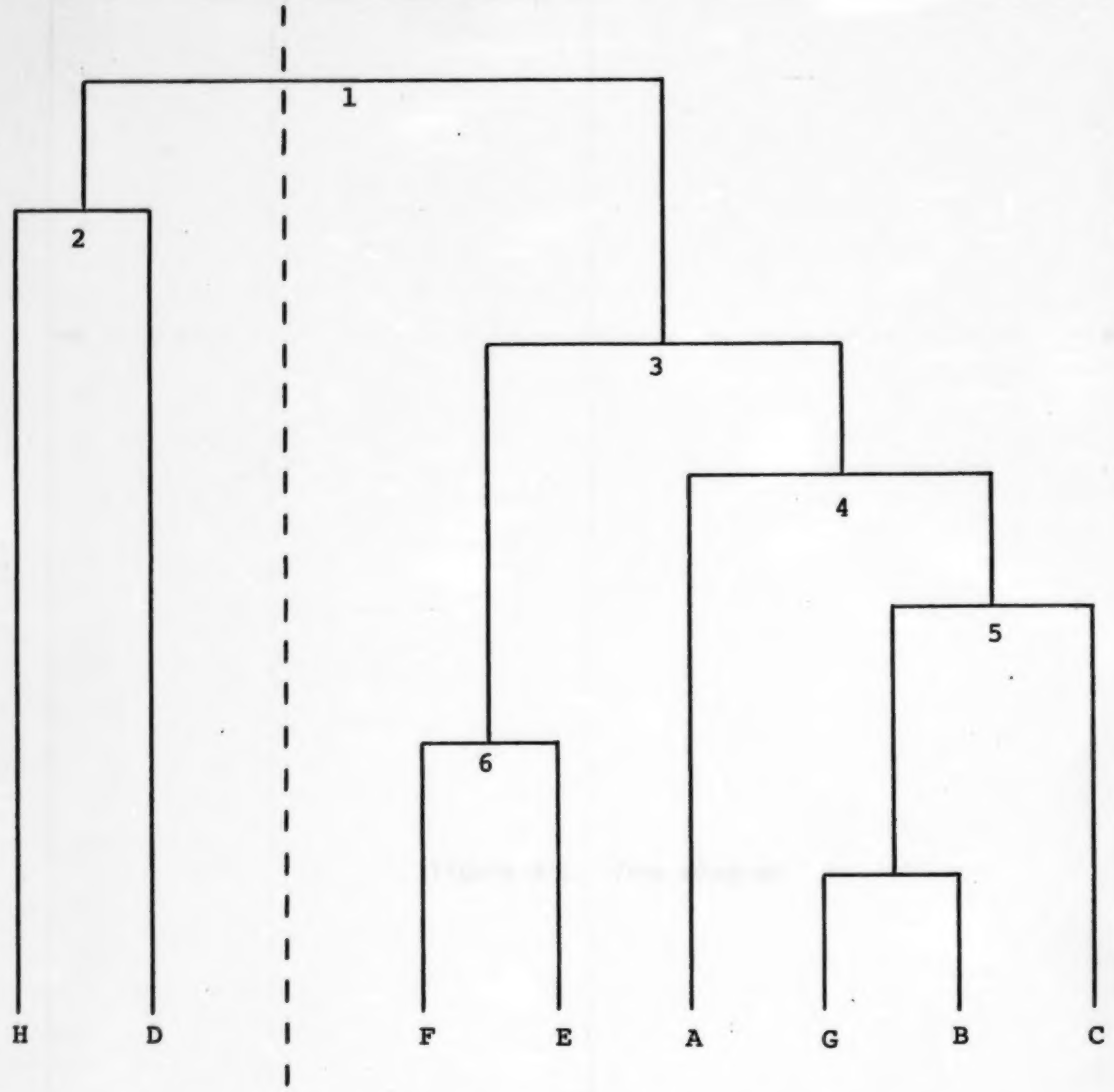

1
2
3
4
5
6
H
D
F
E
A
G
B
C

Figure 3-1. Tree Diagram

3/8''

TRATIONS

PAGE NUMBER

EPA-287 (Cin.)
(4-76)

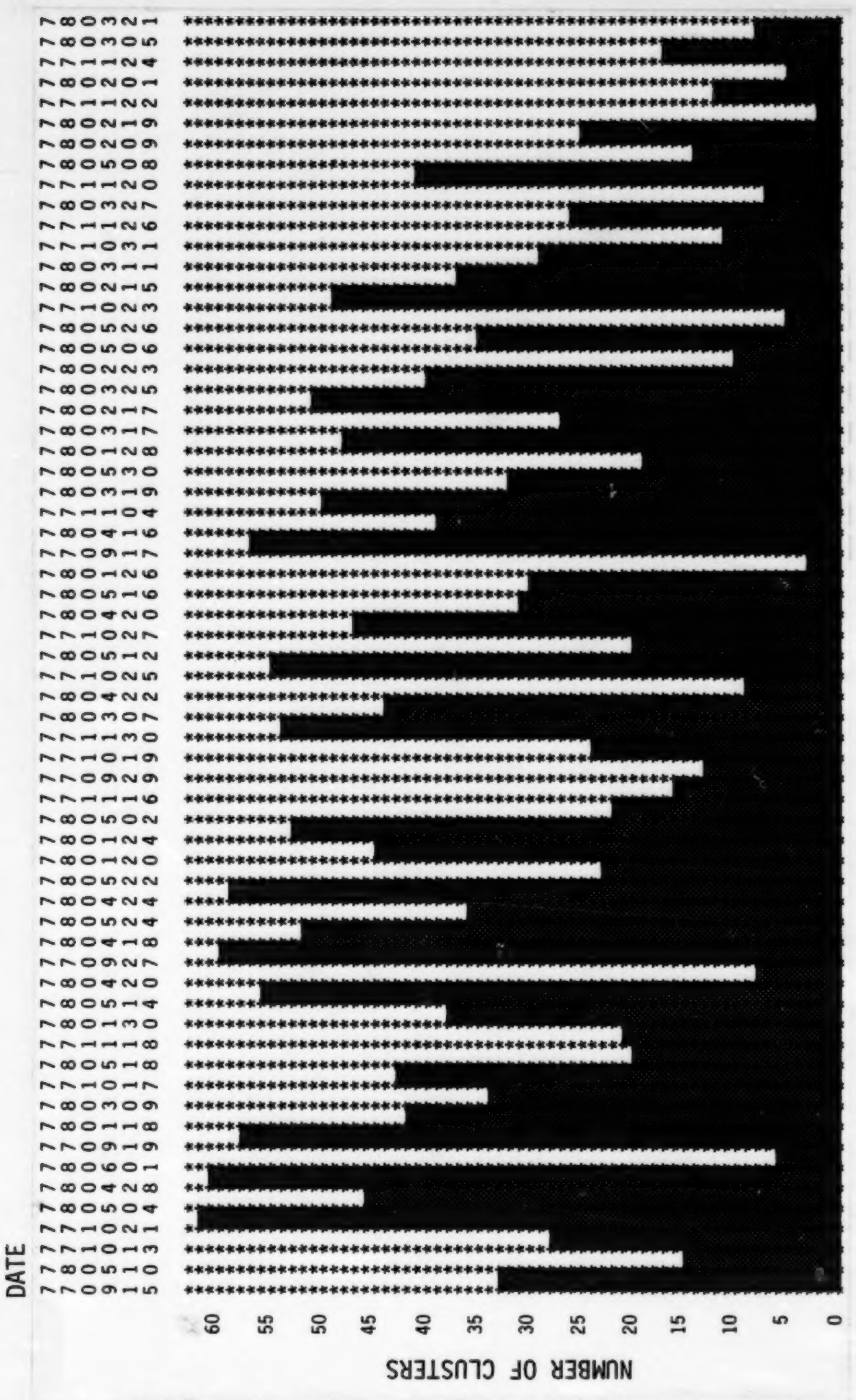


Figure 3-2. Cluster map.

P13

asterisks is a number equivalent to the total number of data points. From the top to the bottom of the map, each new line represents the joining together of two groups. In the computer print-out, when two groups are joined, the space between the asterisks that represent their respective identifiers is filled with asterisks. Here in Figure 3-2, the cluster map, the asterisks and the intermediate asterisks indicating joining have been shaded over to better show the groupings. In a computer print-out, the shaded portions will not be shown but the intermediate asterisks will be seen. The number printed beside each row represents the number of groups at that level. The map continues until all identifiers are joined in one group. The identifiers are arranged at the top so that adjacent points can be joined at some level of grouping.

Another table summarizes the statistics for each level of grouping from one to N that are shown in Figure 3-2.

3.1.4 MIKCA Algorithm

MIKCA[4] (Multivariate Iterative K-means Cluster Analysis) has an advantage over the SAS; it does not impose a hierarchical structure on the data--that is, data points can be switched from one group to another before a final grouping is decided for a specified number of groups. However, the increased number of mathematical computations makes the program runs quite expensive.

The user must decide on the number of groups of interest to enter into the program and pick one of five possible criteria--one of three distance functions, and whether or not to standardize the data.

The criteria options are all measures of the size of the within-groups, sum-of-squares matrix W. This matrix is pooled within groups of squares for each of the groups. Homogeneity of these pooled groups makes it possible to have a small within-groups sum-of-squares. The five criteria are the (1) trace of W; (2) determinant of W; (3) trace of $W^{-1}b$; (4) Bayes modification of the determinant of W; and (5) largest root of the characteristic equation

$$b - \lambda W = 0$$

where b is the between-groups sum-of-squares.

A clustering criterion that is sensitive to disparate group sizes can be useful for identification of outliers.[5] The Bayes modification of the determinant of the within-group sum-of-squares has this potential.[2] The minimization of the unmodified determinant of the within-group sum-of-squares has a tendency to split into equal-sized groups which could "hide" outliers. This will also be very sensitive to singularities and to high correlations between variates.

The optional measures of distance in the program are the (1) Euclidean distance, (2) the weighted Euclidean distance (the weights being the inverse of the variance of each variable), and (3) the Mahalanobis distance (indication of the variance and covariance of the variables). Preference would be the Mahalanobis distance if it were not that the desired variance-covariance matrix is the average (or pooled) covariance matrix for the number of groups for which a solution is desired. This average is only available after the cluster solution is available.

Before the analysis, standardization of the variables--namely, subtraction of the total sample mean and division by the total sample standard deviation of each variable for each observation--is optional. This standardization, which centers the data and scales the variables is an option in MIKCA and the SAS program.

The MIKCA program proceeds by randomly assigning a single data point to each group as an initial choice for each cluster center. As each additional data point is considered, a distance function is calculated, and the observation is assigned to the "closest" cluster; after each assignment, the cluster centers are recomputed, based on an average of the points in the group; finally after all data have been assigned, the value of the criterion function is calculated. This three-step procedure is repeated 10 times before the group with the best criterion value is chosen as the initial cluster solution. Each observation is again considered in order of input and placed in the group that it is closest to. Closest in this step refers to the distance function specified by the user. After all the observations have been placed in respective groups, the criterion value is computed. This process is repeated if the criterion value is decreasing; otherwise, the individual switching heuristic is begun.

The individual switching heuristic now assigns observations to clusters based on the criterion value, rather than a distance function. The program considers a single cluster. It decides if switching any observation in that cluster to another group would improve the criterion value. (Because much computer time can be spent in this process, the user has the option of deciding how many switches per cluster will be allowed; but maximum switching is recommended.) Finally, after all the groups are considered for switching, the program is finished and prints the groups, listing the identifiers in each as well as the cluster centers.

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-10

P15

Presently, the program can handle 500 observations and up to 20 variables. The number of clusters cannot exceed 15. The program can be adapted for larger data sets; however, computer time can become prohibitive if any or all of the above are allowed to increase.

3.1.5 NORMIX Algorithm

The title of this program is formed out of the letters in the words describing the model for this approach: multivariate normal mixtures.[6] Each group is modeled by a multivariate normal distribution; the entire sample is considered as a random sample from a mixture of such components.

The approach is to use maximum likelihood to estimate the parameters in the mixture: a vector of means for each group, a common or separate covariance matrix for the groups and the mixture parameters, which describe the proportions of the total sample from each group. From the estimated parameters, the observations are assigned to the components in the mixture based upon the estimated density height for each component, scaled by its estimated proportional contribution to the total mixture. The assignment of each observation is to the greatest estimated density height, scaled by its estimated contribution to the mixture.

A NORMIX program is available from Wolfe.[6] The most current version of this program was requested and received by the Department of Biostatistics, University of North Carolina at Chapel Hill. For further discussion of the inter-relation of the criteria of MIKCA and NORMIX, see Reference 2.

3.2 ANALYSIS OF 3-7 PM DAILY AVERAGE DATA

Initially SAS was used on the daily average data so a cluster plot (Figure 3-2) could be examined along with the summary statistics of groups 1 to 63. From this run, it was hoped that outliers could be identified and decisions could be made on how many groups should be considered in more depth. Outliers are the observations that resist the longest before joining a group. There were five observations that seemed to fit this description. Their dates and variable values are given in Table 3-1.

The last single observation to join any group was the average for 3/21/78. The observation had the lowest traffic count, a wind direction down the highway, and a windspeed slightly higher than average. Intuitively, it would seem that all of these factors would yield a low CO value. On the contrary, the value of CO (8.8 ppm) was higher than the overall average (7.9).

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

The other values that were suspected of being outliers had less obvious anomalies than these five. The most distinguishing feature of the outlier group as a whole was the fact that they were the observations that had the lowest traffic counts and the highest CO values; which could be explained by a traffic jam. That is, a large number of cars moving at a low speed would result in a low traffic count. The effect of a lot of vehicles sitting or moving slowly could cause high CO readings at the monitor.

TABLE 3-1. POSSIBLE OUTLIER OBSERVATIONS

	Variables Observed						
Date	Suspended particles (μg/m³)	Lead (μg/m³)	CO (ppm)	Wind-speed (mph)	Wind direction	Temp (°F)	Traffic count (10³)
11/22/77	147.0	7.1	13.2	1.8	280	64.6	10.2
2/01/78	115.0	6.7	7.9	4.2	244	61.4	10.5
11/24/77	108.0	11.5	5.3	2.7	232	74.2	9.8
3/05/78	88.0	9.6	4.4	5.1	226	63.4	9.5
3/21/78	83.0	5.5	8.8	5.8	162	63.2	9.1
Averages	108.2	8.1	7.9	3.9	229	65.4	9.8

There is a relationship between the density of cars and the speed of traffic. They both increase up to a point where the density becomes an impedence to speed and then speed drops off as density increases (see Figure 3-3).

In some cases, a group was a single observation which tended to highlight the outliers but was not helpful in determining the interrelations of that observation. Thus, a cluster of eight groups was chosen to more closely study which observations were in each group and what that might mean as far as interrelationships among the variables. Even for a cluster of eight groups, the number per group was so small that the differences between the groups was difficult to decipher. For eight groups, the hierarchical scheme resulted in the breakdown of the data as shown in Table 3-2. The five suspected outliers fell in groups 7 and 8 with two and three members each. Four of the six remaining groups had nine or fewer members, and groups 3 and 4 had 12 or more members. This disparity in group sizes suggests that a smaller number of groups should be considered to get more per group.

EPA-287 (Cin.)
(4-76)

9-12

BEGIN
FIRST
LINE OF
TEXT
HERE
TYPING GUIDE SHEET
CENTER
OF PAGE
TOP OF
IMAGE
AREA
DROPPED
HEAD;
BEGIN
SECTIONS

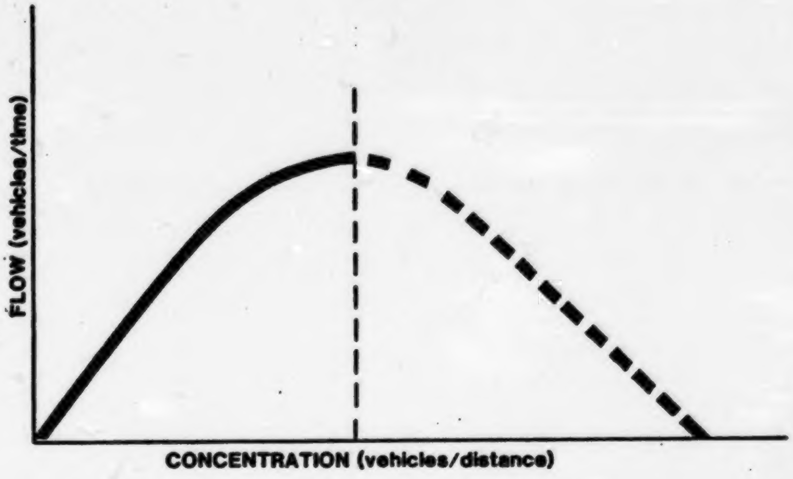


Figure 3-3. General flow-density relationship.

TTOM OF
AGE AREA;
UTSIDE
DIMENSION
FOR TABLES
AND ILLUS-
TRATIONS
LAST LINE
OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLE 3-2. CLUSTER CENTERS OF EIGHT GROUPS: 3-7 PM DAILY AVERAGE DATA

	Group Center						
Group no. and no. per group	Suspended particles ($\mu g/m^3$)	Lead ($\mu g/m^3$)	CO (ppm)	Wind-speed (mph)	Wind direction	Temp (°F)	Traffic count (10^3)
1-07	153.0	9.9	6.8	5.4	225	67.8	13.6
2-09	138.3	9.1	7.2	5.1	218	67.4	12.6
3-20	168.3	9.9	6.8	4.7	217	68.8	13.0
4-12	122.7	7.6	7.3	5.4	230	68.9	12.0
5-06	137.8	8.8	7.5	4.7	223	66.9	11.5
6-04	124.5	8.1	5.8	6.5	225	66.6	11.0
7-02	131.0	6.9	10.5	2.9	262	63.0	10.4
8-03	93.0	8.9	6.1	4.5	207	66.9	9.5

6-1/2"

variation in temperature from 63.0° to 68.9° seemed to have little influence on any of the pollutants; this was no surprise since the temperature in this area varies so little day by day that it was not expected to be an important predictor of pollutant levels. Suspended particles and lead levels had a direct relationship with traffic counts.

Wind directions for groups 1-6 were within a range that could be considered blowing across the highway towards the monitoring site. The wind directions of groups 7 and 8 (262 and 206) were not quite as in line with the highway and monitoring site as other values; however, the CO level for group 7 was the highest and the accompanying traffic count was the second lowest. The group with the lowest traffic count, group 8, also had a low CO average.

The number of clusters was reduced from eight to four (Table 3-3). The four-group cluster combined the two outlier groups into one group; this clustering again showed the relationships noted earlier with the eight-group cluster. The first three groups of the eight-group cluster were combined with one group. Group 4 remained the same, and the remaining 15 observations were split into two groups of 10 and 5 members each. Another point was that each group seemed to be split on the basis of traffic counts; there were no overlapping traffic counts between groups, but the same was not the case for any of the other variables. This, of course, suggests that traffic is the main determining factor in these data.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLE 3-3. CLUSTER CENTER OF FOUR GROUPS: 3-7 PM DAILY AVERAGE DATA

Group no. and no. per group	Center of Group: Suspended particles (μg/m³)	Lead (μg/m³)	CO (ppm)	Wind-speed (mph)	Wind direction	Temp (°F)	Traffic count (10³)
1-36	157.8	9.7	6.9	4.9	219	68.2	13.0
2-12	122.7	7.6	7.3	5.4	230	68.9	12.0
3-10	132.5	8.5	6.9	5.4	223	66.8	11.3
4-05	108.2	8.1	7.9	3.9	228	65.4	9.8

The MIKCA program was run with the daily data to compare with data on the four-group the SAS procedure output. An initial run was made with the following inputs: four groups, nonstandardization of the data, Euclidean distance, and minimization of the determinant of W for the criterion. The results were similar to those of SAS. The MIKCA listed the observations associated with each cluster. The corresponding months and days of the week are given in Table 3-4.

The observations in cluster 1 came from the weekdays--that is, the days with the highest traffic counts. This is to be expected since the traffic volume between 3 and 7 PM is normally the highest during the week. Each group's observations were spread over the entire study period, indicating that seasonality plays an undiscernible role in these data. In comparing the groups of MIKCA with those of SAS, one group was identical; this was the group of outliers. The other three groups were similar. Table 3-5 compares the groups for the two algorithms. On comparison of the two techniques, all four groups were reasonably similar. The group one comparison showed slight dissimilarities so a three-group solution was obtained. Results were more divergent on the three-group (Table 3-6) than on the four-group cluster. The comparison of three-group clusters showed that the MIKCA algorithm had made several finer adjustments because single observations had been switched to other groups.

3.3 ANALYSIS OF FOUR 1-HOUR PERIODS PER DAY DATA

Because there were so few daily average data, four 1-hour periods in each day were analyzed with the MIKCA algorithm. The resulting data set had 523

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TYPING GUIDE SHEET
BEGIN FIRST LINE OF TEXT HERE
CENTER OF PAGE
TOP OF IMAGE AREA
DROPPED HEAD; BEGIN SECTIONS HERE
6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLE 3-4. FREQUENCY TABLES BY DAY OF WEEK AND MONTH FOR MIKCA RESULTS FOR FOUR GROUPS 3-7 PM DAILY AVERAGE DATA

	Suspended particles (μg/m³)	Lead (μg/m³)	CO (ppm)	Wind-speed (mph)	Wind direction	Temp (°F)	Traffic count
CLUSTER 1:							
Center	176.6	10.03	6.94	5.19	208.15	69.1	13,223
CLUSTER 2:							
Center	108.2	8.08	7.93	3.90	228.8	65.4	9,815
CLUSTER 3:							
Center	126.26	8.14	7.53	5.04	225.7	68.1	11,614
CLUSTER 4:							
Center	134.1	8.93	6.46	5.02	233.4	67.2	12,695

CLUSTER 1:

Day of Week	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Total
	2	3	4	7	4			20

Month	Sept	Oct	Nov	Apr	May	June	Total
	2	5	2	3	7	1	20

CLUSTER 2:

Day of Week	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Total
		2	1	1			1	5

Month	Nov	Feb	Mar	Total
	2	1	2	5

CLUSTER 3:

Day of Week	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Total
	3	1	2	1	3	5	4	19

Month	Sept	Oct	Nov	Jan	Feb	Mar	May	Total
	1	2	3	1	5	5	2	19

CLUSTER 4:

Day of Week	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Total
	3	4	2	3	2	3	2	19

Month	Sept	Oct	Nov	Jan	Mar	Apr	May	Total
	2	1	1	5	2	3	5	19

TABLE 3-5. COMPARISON OF MIKCA AND SAS CLUSTER CENTERS FOR FOUR GROUPS: 3-7 PM DAILY AVERAGE DATA

		Center of Group						
Algorithm	Group No. and No. per Group	Suspended particles (μg/m³)	Lead (μg/m³)	CO (ppm)	Wind-speed (mph)	Wind direction	Temp (°F)	Traffic count (10³)
SAS	1-36	157.8	9.7	6.9	4.9	219	68.2	13.0
MIKCA	1-20	176.6	10.0	6.9	5.2	208	69.1	13.2
SAS	2-05	108.2	8.1	7.9	3.9	229	65.4	9.8
MIKCA	2-05	108.2	8.1	7.9	3.9	229	65.4	9.8
SAS	3-10	122.7	7.6	7.3	5.4	230	68.9	12.0
MIKCA	3-19	126.3	8.1	7.5	5.0	225	68.0	11.6
SAS	4-10	132.5	8.5	6.9	5.4	224	66.8	11.3
MIKCA	4-19	134.1	8.9	6.5	5.0	233	67.2	12.7

6-1/2"
9-1/8"

TABLE 3-6. COMPARISON OF MIKCA AND SAS CLUSTER CENTERS FOR THREE GROUPS: 3-7 PM DAILY AVERAGE DATA

		Center of Group						
Algorithm	Group No. and No. per Group	Suspended particles (μg/m³)	Lead (μg/m³)	CO (ppm)	Wind-speed (mph)	Wind direction	Temp (°F)	Traffic count (10³)
MIKCA	1-09	135.9	8.3	6.7	4.4	238	67.7	10.4
SAS	1-22	127.1	8.0	7.1	5.4	227	67.9	11.7
MIKCA	3-32	161.2	9.8	7.1	4.6	218	67.6	13.1
SAS	2-36	157.8	9.7	6.9	4.9	219	68.2	13.0
MIKCA	3-22	120.7	8.1	7.1	5.8	224	68.3	12.0
SAS	3-05	108.2	8.1	7.9	3.9	229	65.4	9.8

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

observations; deleting the first 24 made it fall within the 500-observations limitation of MIKCA. The SAS procedure could not be used because of its limitation of 250 observations.

The four 1-hour periods were 2 AM, 8 AM, 2 PM, and 5 PM--2 hours of rush hour traffic, morning and afternoon; and 1 hour of very low traffic, 2 AM; and 1 hour of moderate midday traffic. These hours were thought to cover the full range of conditions, including diurnal variations.

To begin, a cluster of eight groups was selected. If a seasonal/diurnal effect should become evident, eight groups would be likely to allow for each of the 4-hour periods to split into one of two seasons.

As in the previous analysis on daily data, seasonality was not a factor. Each group had observations from throughout the study period. The groups did split on the 4 hours of the day. The 3 AM observations were segregated totally onto a single group, averaging only 828 cars and 1.2 ppm of CO. The other groups had the daytime observations mixed (Table 3-7). As the traffic counts and CO levels increased, the observations shifted from 3 AM to 8 AM to 2 PM to 5 PM, except in group 8 which had over half its observations at 8 AM. In this data set, CO was positively correlated with traffic counts, an intuitive relationship. Windspeed and direction as well as temperature seemed to be of little significance. Groups 3 and 4 showed the wind to be of little significance, and both groups showed the wind to be blowing virtually down the highway; however, group 2 has an average wind direction towards the highway from the monitoring site. These three groups got most of their observations from the 8 AM data. The CO levels seemed to increase as each day progressed. This could be accounted for by the increased traffic flow and by the prevailing wind direction.

A four-group cluster was run next, since seasonality seemed to have no effect. These results are summarized in Table 3-8. The groups seemed to be split more obviously by hours. Again high traffic counts pointed to high CO levels. Temperature and hour of the day were related, but temperature alone gave no new information on CO. The four-group analysis was run to minimize Bayes factor 1 of determinant W (see Reference 5). The optimal four-group results in Table 3-9 were similar to those in Table 3-8.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

TABLE 3-7. CLUSTER CENTERS FOR EIGHT-GROUP ANALYSIS FOR FOUR 1-HOUR PERIODS PER DAY DATA

					Center of Group				
Group No. and No. per group	Hour of Day AM 3	Hour of Day AM 8	Hour of Day PM 2	Hour of Day PM 5	CO (ppm)	Wind direction	Wind-speed (mph)	Temp (°F)	Traffic count (10^3)
1-123	121	0	1	1	1.2	263	2.2	57.8	0.8
2-14	0	12	1	1	3.0	83	3.1	66.2	4.2
3-10	0	8	1	1	3.2	162	3.5	68.9	5.7
4-20	0	17	2	1	3.8	147	2.9	67.7	8.1
5-24	0	6	12	5	5.3	205	6.0	69.4	10.0
6-116	0	6	80	30	6.5	230	5.3	71.0	11.8
7-77	0	15	25	37	6.9	221	4.6	68.5	12.9
8-115	0	59	4	52	6.2	190	3.6	66.3	14.0

9-1/8"

TABLE 3-8. CLUSTER CENTER FOR FOUR-GROUP ANALYSIS: CRITERION TO MINIMIZE DETERMINANT W FOR FOUR 1-HOUR PERIODS PER DAY DATA

					Center of Group				
Group No. and No. per group	Hour of Day AM 3	Hour of Day AM 8	Hour of Day PM 2	Hour of Day PM 5	CO (ppm)	Wind direction	Wind-speed (mph)	Temp (°F)	Traffic count (10^3)
1-125	121	0	1	2	1.2	260	2.8	57.8	0.8
2-44	0	38	4	2	3.5	127	3.0	68.2	6.6
3-164	0	18	107	39	4.9	234	6.2	71.7	12.0
4-166	0	66	15	85	8.0	193	3.0	65.6	13.4

BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
EPA-287 (Cin.)
(4-76)

9-19

TABLE 3-9. CLUSTER CENTER FOR FOUR-GROUP ANALYSIS: CRITERION TO MINIMIZE BAYES FACTOR 1 OF DETERMINANT W FOR FOUR 1-HOUR PERIODS PER DAY DATA

Group No. and No. per group	Hour of Day AM 3	8	PM 2	5	Center of Group CO (ppm)	Wind direction	Wind-speed (mph)	Temp (°F)	Traffic count (10^3)
1-125	121	0	1	8	1.2	260	2.8	57.8	0.8
2-43	0	38	3	2	3.4	124	2.9	67.9	6.5
3-159	0	17	105	37	4.9	233	6.3	72.1	11.9
4-172	0	67	18	87	7.8	195	3.1	65.6	13.4

3.4 APPLICATIONS OF CLUSTER ANALYSIS

There are several references illustrating the application of cluster analysis to problems in air pollution.[10-14] In almost all of these examples, the cluster analysis was used in conjunction with factor analysis and other multivariate techniques. The typical problem considered was one in which the concentrations of several chemical elements (one example also involved several site locations) were measured on many pollutant samples. The matrix of simple correlation coefficients was obtained for the elements. The similarity between elements was thus determined by r, the pair-wise correlation, or a measure of dissimilarity (distance) by 1 - r. Dendograms and nonlinear mappings were given to illustrate the groupings of the chemical elements. These clusters were then associated with various sources of the pollutant--for example, soils on which the crucial elements were highly correlated, automotive sources, combustion processes, mining activity. In summary, the cluster analysis was helpful in a preliminary analysis of these data; however, in almost all cases the analyses were followed by a more objective procedure--factor analysis for identifying the factors.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-20

4.0 MULTIPLE REGRESSION ANALYSIS

It was not the purpose of this study to predict the concentration of CO observed at site 5 as a function of wind direction and windspeed, traffic count, and temperature. However, it was decided that a small effort could be made to obtain these relationships and to compare them with those of the cluster analysis. It might also be argued that the regression approach would be a more natural way to analyze these particular data.

It is obvious that the CO concentrations should be a function of each of the variables; however, it is not known how much of the variation in the CO concentrations would be explained by the variables, singly or in combination.

Different regression analyses were performed on the hourly data to find the best way of handling the wind direction and windspeed in combination. The first model (or relationship) for the predicted mean CO concentration is as follows:

$$\text{CO (predicted)} = b_0 + b_1\,(\text{windspeed}) + b_2\,(\text{WS}) + b_3\,(\text{wind direction}) + b_4\,(\text{temperature}) + b_5\,(\text{traffic count}),$$

where

CO = the predicted mean CO concentration, ppm,
Speed = windspeed, mph units,
WS = wind direction x windspeed,
Wind direction = 0, if direction $>325°$ or is $<125°$
1, if $125° <$ direction $<325°$.

That is, if the wind direction is toward site 5, independent of the angle with respect to the road, then the variable takes on the value 1; otherwise the value is 0.

Temperature = ambient temperature in °F,
Traffic count = total traffic volume--that is, number of vehicles per hour for all traffic lanes in both directions.

The prediction equation resulting from the fitting of this model by the method of least squares is:

PAGE NUMBER
EPA-287 (Cin.)
(4-76)

$$\text{CO (predicted)} = -0.303 - 0.425\ \text{(speed)} + 0.217\ \text{(WS)} + 0.945\ \text{(wind direction)} + 0.030\ \text{(temperature)} + 0.000369\ \text{(traffic count)}$$

Each of the variables makes a (statistically) significant contribution to the prediction of the CO concentration. The square of the multiple correlation coefficient is 0.468, which means that the linear prediction relationship explains 47% of the total variation in the CO concentrations. The standard error of the estimate is 2.31 ppm. Thus, it was of further interest to investigate another nonlinear model.[8]

The second model investigated was:

$$\text{CO (predicted)} \cong b_0 + b_1\ \text{(temperature)} + b_2\ \text{(traffic count)} + b_3\ \text{(WSI)},$$

where

WSI = the product variable, windspeed × Cos (direction-235°).

This WSI combination variable is the component of windspeed in the direction normal to the flow of traffic. For this three-variable prediction, the results are:

$$\text{CO (predicted)} \cong -0.939 + 0.037\ \text{(temperature)} + 0.000430\ \text{(traffic count)} + 0.229\ \text{(WSI)}$$

The square sum of the multiple correlation coefficient is 0.405, and the standard error of estimate is 2.43 ppm CO, compared to 0.468 and 2.31, respectively, from the five-variable model.

To obtain a relative ranking of importance of the five variables and/or the models with fewer variables, a stepwise regression procedure was used to add one variable at a time. The resulting order of addition is as follows:

1. Traffic count,
2. Wind direction (0, 1),
3. Speed (wind),
4. WS (wind direction × windspeed), and
5. Temperature.

BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-22

There appears to be no surprises here in terms of the order, with traffic alone, $R^2 = 0.38$, and after the third variable was added, $R^2 = 0.458$, compared with the 0.468 R^2 value for all five variables.

A stepwise regression for the three-variable model yielded identical ordering (to that of the five-variable model) of the importance of the variables; that is, the order of entry in the stepwise regression was:

1. Traffic count,
2. WSI, and
3. Temperature.

6-1/2"
9-1/8"
BEGIN LAST LINE OF TEXT
BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-23

5.0 CONCLUSIONS

The two clustering algorithms used were each useful for different purposes. The SAS gave the cluster map for deciding on a preliminary estimate of the number of groups and for finding outliers. The main disadvantage of SAS was the small size (under 250 observations) of the input data set.

The MIKCA permitted more generalized clustering and more flexibility in several of its criteria than SAS. MIKCA is also more sensitive than SAS to outliers and avoids the strictness of a hierarchical structure.

It was possible using cluster analysis to identify variable relationships; a data set could easily be envisioned in which no inner relationships of the variables were previously known, so clustering was helpful for preliminary analyses. However, the relationships were more obvious with other multivariate techniques.

A thorough comparison of the various programs and options on distance, standardization, criteria, and number of groups would require considerable amounts of additional effort and expense. A more thorough effort could include:

1. An investigation of the effect, if any, standardization would have on the data (a relative standardization, namely $(x - \bar{x})/s$) would be investigated to better scale the data;
2. A comparison of the solutions for various numbers of groups with a fixed criterion;
3. A comparison of cluster solutions from several additional clustering programs, including NORMIX and those in the BMD* series; and
4. Observations of the impact of each variable on groupings by making clusters without certain variables.
5. An investigation into traffic density and its effect on pollutant levels.

---

*HBMD (Biomedical Computer Programs) is a software package that (like SAS) incorporates many statistical algorithms among which is a clustering procedure.

PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-24

6.0 REFERENCES

1. The Los Angeles Catalyst Study Symposium. EPA-600/4-77-034, June 1977.

2. Marriott, F. H. Cluster Analysis and Related Problems, The Interpretation of Multiple Observations, Academic Press, 1974.

3. Barr, A. J., et al. A User's Guide to SAS-76. Spanks Press, Raleigh, North Carolina, 1976.

4. McRae, D. J. MIKCA: A FORTRAN IV Iterative K-Means Cluster Analysis Program. CTB/McGraw-Hill, Del Monte Research Park, Monterey, California, September 1970. Revised June 1973 by M. J. Symons.

5. Symons, M. J. Bayes Modification of Some Clustering Criteria. Institute of Statistics Memo Series No. 880. Revised October 1977.

6. Wolfe, J.H. NORMIX: Computation Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. Research Memo, SRM 68-2. U.S. Naval Personnel Research Activity, San Diego, California.

7. Hartigan, J. A. Clustering Algorithms. John Wiley and Sons, Inc., 1975.

8. Tiao, G. C., M. S. Phadke, and G. E. Box. Some Empirical Models for the Los Angeles Photochemical Smog Data. Journal of the Air Pollution Control Association, 26:485-490, May 1976.

9. Everitt, B.S. Unresolved Problems in Cluster Analysis. Biometrics, 35:169-181, March 1979.

10. Draft Report on the Digest on Ambient Particulate Analysis and Assessment Methods (to be published) for the Workshop on Ambient Particulate Analysis and Assessment Methods.

11. University Analytical Center, University of Arizona. Identification and Analysis of Urban Air Quality Patterns. Prepared for Electric Power Research Institute, Palo Alto, California, December 1977.

12. Fordyce, J.S., R. B. King, and H. E. Neustadter. Elemental Composition of Airborne Particulates and Source Identification: Data Analysis Techniques. J. Air Poll. Control Assoc., 26:1079-1084, November 1976.

13. Gladney, E.S., et al. The Use of Multivariate Analysis to Identify Sources of Selected Elements in Boston Urban Aerosol. Atm. Environ., 10:1015-1025, May 1976.

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS
BEGIN LAST LINE OF TEXT
3/8"
PAGE NUMBER
EPA-287 (Cin.)
(4-76)

9-25

P30

TYPING GUIDE SHEET

BEGIN FIRST LINE OF TEXT HERE

CENTER OF PAGE

TOP OF IMAGE AREA

DROPPED HEAD; BEGIN SECTIONS HERE

6-1/2"

9-1/8"

BEGIN LAST LINE OF TEXT

BOTTOM OF IMAGE AREA; OUTSIDE DIMENSION FOR TABLES AND ILLUSTRATIONS

3/8"

PAGE NUMBER

EPA-287 (Cin.)
(4-76)

14. Kowalski B. R. and C. F. Bender. Pattern Recognition: A Powerful Approach to Interpreting Chemical Data. Journal of the American Chemical Society, 94:16, August 1972.

9-26